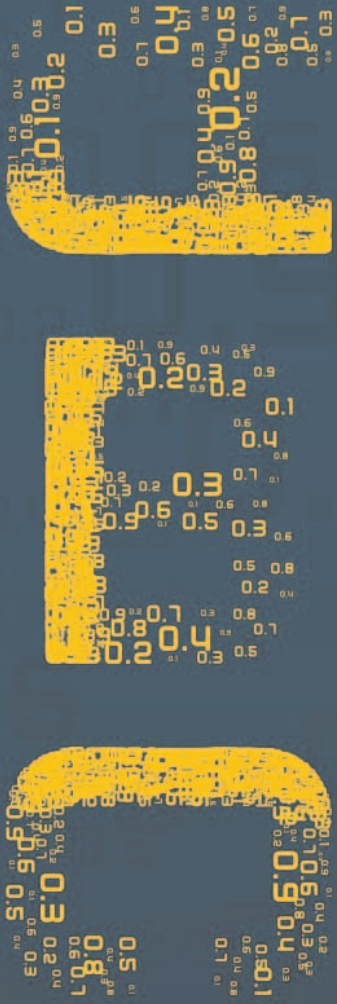
CENEVAL®



El ABC
de la evaluación educativa
(Measuring Up)

DANIEL KORETZ

El ABC de la evaluación educativa

(*Measuring Up*)

Daniel Koretz

CENEVAL®

Centro Nacional de Evaluación para la Educación Superior, A.C.

MÉXICO ■ 2010

Título original:
Measuring Up
What Educational Testing Really Tells Us
Daniel Koretz

El ABC de la evaluación educativa
(Measuring Up)

Traducción:
María Elena Ortiz Salinas

Revisión técnica:
Rafael Vidal Uribe
Eduardo Andere Martínez

© 2008 by the Presidente and Fellows of Harvard College

D.R. © 2010, Centro Nacional de Evaluación
para la Educación Superior, A.C. (Ceneval)
Av. Camino al Desierto de los Leones 19,
Col. San Ángel, Deleg. Álvaro Obregón,
C.P. 01000, México, D.F.
www.ceneval.edu.mx

Diseño de portada:
Alvaro Edel Reynoso Castañeda

Diseño y formación de interiores:
Mónica Cortés Genis

Se prohíbe la reproducción total o parcial de esta
obra –incluido el diseño tipográfico y de portada–,
sea cual fuere el medio, electrónico o mecánico,
sin el consentimiento por escrito del editor.

Agosto de 2010

ISBN: 978-607-7704-01-0

Impreso en México ▪ Printed in Mexico

Contenido

- Prólogo 5

1. Si sólo fuera tan simple 9
2. ¿Qué es una prueba? 21
3. Lo que medimos: ¿qué tan buena es la muestra? 43
4. La evolución de la evaluación en Estados Unidos 55
5. ¿Qué nos dicen las calificaciones de las pruebas sobre los niños estadounidenses? 89
6. ¿Qué influye en los resultados de las pruebas? (o cómo no escoger una escuela) 135
7. Error y confiabilidad: ¿qué tanto no sabemos de lo que estamos hablando? 169
8. Informe sobre desempeño: estándares y escalas 209
9. Validez 251
10. Resultados inflados de las pruebas 275
11. Impacto adverso y sesgo 303
12. Evaluación de estudiantes con necesidades especiales 327
13. Usos razonables de las pruebas 367

- Notas 387
- Reconocimientos 397
- Índice temático 399

El ABC
de la
evaluación educativa

Prólogo

En casi todo Estados Unidos es difícil exagerar la importancia de la evaluación educativa: ha sido enorme su impacto en la práctica educativa y es un factor relevante en las decisiones de innumerables familias respecto a dónde vivir y si deben utilizar las escuelas oficiales. Las pruebas tienen también una influencia poderosa en el debate público acerca de temas sociales como la competitividad económica, la inmigración y las desigualdades raciales y étnicas. Además, debido a su sencillez y sentido común, las pruebas de logro parecen ser dignas de confianza: se pide a los estudiantes que realicen ciertas tareas, se ve cómo las realizan y de este modo se juzga qué tan exitosos son ellos y sus escuelas.

Sin embargo, esta aparente simplicidad es engañosa. La evaluación del logro es una empresa muy compleja y, por ende, es común que se malentienda y se abuse de los resultados de sus pruebas; las consecuencias pueden ser graves precisamente por la importancia que se confiere en nuestra sociedad a las calificaciones obtenidas en las pruebas. Entonces, la intención de este libro es ayudar a los lectores a entender las complejidades inherentes a la evaluación, evitar los errores comunes, hacer una interpretación razonable de las calificaciones y un uso productivo de las pruebas.

La evaluación se ha convertido en tema de acalorada controversia y en los años recientes se ha publicado un gran número de polémicas, con argumentos en favor y en contra. No es el caso de este libro, aunque en él se hace amplia referencia a las controversias

políticas relacionadas con la evaluación y no se rehúye la explicación de las posturas que, en mi opinión, están mejor sustentadas por la evidencia. Por ejemplo, la investigación muestra de manera sistemática que la evaluación de alto impacto (la que hace que la gente se responsabilice de mejorar las calificaciones) puede producir una enorme inflación de los resultados, y en un capítulo posterior describo parte de esa evidencia. No obstante, mi propósito no es hacer una apología de una u otra postura. Más bien, pretendo aclarar las virtudes y las limitaciones de la evaluación del logro. Si bien puede cometerse un abuso atroz cuando se usan los resultados sin considerar esas limitaciones, al emplearse con cuidado arrojan información valiosa que no puede obtenerse de otras fuentes. Para extraer lo bueno y evitar lo malo es necesario conocer algo sobre las dificultades de la evaluación, de la misma manera que se requiere de conocimiento y cautela cuando se consume un medicamento potente y se quieren evitar efectos secundarios peligrosos. Hace más de una década, Don Stewart, en ese momento presidente del Consejo de Universidades (una organización que auspicia el SAT, uno de los exámenes más conocidos y de mayor uso en Estados Unidos), lo planteó de forma clara: dijo que al usar las pruebas es necesario encontrar un equilibrio entre el entusiasmo y el respeto por la evidencia.

Algunas de las complejidades de la evaluación del logro parecen enormes, incluso incomprensibles para la mayoría de los usuarios de los resultados de las pruebas. Puede entenderse el desconcierto del desventurado padre o del director de escuela que descarga el informe técnico del programa de evaluación de su estado y lee, por ejemplo, que los resultados se escalaron usando un modelo logístico de tres parámetros de la teoría de respuesta al ítem.

La buena noticia es que si bien esos intimidantes detalles técnicos son de fundamental importancia para la elaboración de las pruebas y la operación de los programas de evaluación, no son

necesarios para entender los principios de la evaluación y las controversias a que ésta da lugar. En la Escuela de Posgrado en Educación de Harvard cada año se inscribe un gran número de estudiantes de maestría que deben conocer lo básico de la evaluación educativa; necesitan convertirse en consumidores informados de los resultados y en usuarios cuidadosos de las pruebas, pero no cuentan con la formación matemática, el tiempo o la necesidad de la introducción matemática tradicional a la medición. Hace algunos años preparé un curso justo para esos estudiantes, diseñado para hacerles entender los principios de la medición y las controversias actuales en ese campo, como la evaluación de alto impacto y la evaluación de estudiantes con discapacidades. El curso no requiere formación previa en matemáticas y su contenido en ese tema es mínimo. Ese curso dio lugar a este libro.

El libro evita las presentaciones matemáticas, incluso más que el curso que lo propició. No incluye ecuaciones y las gráficas son muy pocas. Sin embargo, no evade el uso de conceptos técnicos, como confiabilidad y error de medición, porque es necesario conocerlos para poder entender y hacer juicios fundamentados acerca de algunos de los problemas que surgen en la evaluación. Así, los lectores encontrarán algunas secciones didácticas diseminadas en el libro, en ocasiones como capítulos separados y otras veces no. He intentado organizar las cosas de modo que los lectores puedan saltarse esas partes si quieren, pero lo hice a regañadientes porque el material es importante y tengo la esperanza de que todos se tomarán el tiempo de asimilarlo.

El lector encontrará también que es útil conocer algunos convencionalismos que seguí al escribir este libro. Los académicos salpican sus textos con citas para apoyar sus afirmaciones y dar a otros colegas el crédito que merecen, pero las citas pueden resultar una molesta distracción para los lectores profanos en la materia. No obstante, es necesario incluir algunas citas, por cuestión de

ética (para dar a los demás el crédito por su trabajo) y para el lector ocasional que desea más detalles. Por lo tanto, presento un pequeño número de citas para dar crédito o hacer referencia a las fuentes, pero las presento como notas finales para sacarlas de la vista. En contraste, los lectores quizá *quieran* ver una nota explicatoria ocasional, por lo que estas se presentan como notas al pie de la página relevante. Los términos técnicos se explican la primera vez que se usan. ■

Capítulo 1
Si sólo fuera tan simple

Hace unos años recibí una llamada de una desconocida que estaba a punto de mudarse a mi distrito escolar y quería que la ayudara a identificar buenas escuelas. Supuso que yo podía hacerlo debido a mi trabajo. Tomé su inquietud más en serio de lo que ella deseaba y le dije en pocas palabras lo que yo buscaría, no sólo como experto en evaluación e investigación educativa sino también como padre de dos niños en edad escolar y como antiguo maestro de primaria y secundaria. Sugerí que, como primer paso, debía recabar tanta información descriptiva como pudiese para hacerse una idea de las escuelas que querría considerar. En mi lista de información descriptiva las calificaciones obtenidas en las pruebas ocuparían un lugar elevado pero, dependiendo del niño, también le daría importancia a muchas otras cosas como la calidad de los programas musicales o deportivos de la escuela, algunos énfasis especiales del programa, el tamaño de la escuela, la diversidad social, etcétera. Le dije que una vez que hubiese disminuido su lista lo suficiente (este es un distrito escolar muy grande), debería visitar algunas de las escuelas que le parecieran prometedoras. Una visita le permitiría vislumbrar las características de los planteles, incluyendo aquellas que podrían contar para sus calificaciones en la prueba. Le expliqué algunas de las cosas que yo consideré cuando supervisaba escuelas y aulas para mis propios hijos; por ejemplo: un alto nivel de participación de los estudiantes, explicaciones claras de los maestros antes de que los alumnos iniciaran

las tareas, un nivel de actividad entusiasta cuando fuese apropiado y discusiones animadas entre los educandos. Al contar con las observaciones y la información descriptiva, estaría en mejores condiciones de identificar las escuelas adecuadas para sus hijos.

La mujer no estaba contenta. Quería una respuesta sencilla que implicara menos trabajo, o siquiera menos ambigüedad y complejidad. Una respuesta sencilla es tranquilizante, sobre todo cuando están en juego la educación de sus hijos y una gran cantidad de dinero. (Esto fue en Bethesda, Maryland, donde los precios de las viviendas son escandalosamente altos.)

Unas semanas más tarde le mencioné esa conversación a un amigo que en esa época dirigía un importante programa de evaluación. Me contó que él recibía ese tipo de llamadas todo el tiempo y que muy pocos de sus interlocutores quedaban contentos con sus respuestas. Querían algo más sencillo: los nombres de los planteles con las calificaciones más altas, que era lo que esas personas consideraban suficiente para identificar a las mejores escuelas. Me contó que en una conversación perdió la paciencia cuando la mujer que lo llamaba se negaba a aceptar una explicación más razonable, por lo que al final le dijo: "Si todo lo que usted quiere es un promedio elevado de calificaciones, dígale a su agente de bienes raíces que quiere comprar una propiedad en el vecindario de mayor ingreso que pueda manejar. Eso le permitirá obtener el promedio de calificación más alto que pueda permitirse".

La llamada de la mujer a la que así se sugirió comprar casa ostentosa refleja dos ideas erróneas sobre la evaluación del logro: que las calificaciones de un solo examen nos dicen todo lo que necesitamos conocer sobre el logro del estudiante y que esta información nos dice todo lo que necesitamos conocer sobre la calidad de la escuela. En capítulos posteriores explicaremos por qué ninguna de esas suposiciones comunes está justificada. Otra equivocación común es que la evaluación es algo sencillo. Al inicio de su

primer mandato presidencial George W. Bush, uno de cuyos programas clave –«Que ningún niño se quede atrás»– se construyó alrededor de la evaluación, declaró: "Una prueba de comprensión de lectura es una prueba de comprensión de lectura. Y una prueba de matemáticas de cuarto grado –no hay muchas formas de poder engañar una prueba... Es muy sencillo 'normar' los resultados".[1] Independientemente de lo que uno piense del programa «Que ningún niño se quede atrás» (y hay buenos argumentos a favor y en contra de varios de sus elementos), esa afirmación es del todo errónea: es muy sencillo errar el diseño de una prueba e incluso es muy fácil equivocarse en la interpretación de los resultados. Y Bush no es el único que tiene esta visión equivocada. Hace algunos años, la representante de un destacado grupo de empresas se dirigió a una reunión del Comité de Evaluación y Examinación del Consejo Nacional de Investigación, del que yo formaba parte. Se quejó de que sus jefes, algunos de los directivos estadounidenses más importantes que participaron en la reforma educativa, estaban exasperados porque nosotros, los profesionales de la medición, proporcionábamos respuestas mucho más complicadas de lo que ellos deseaban. Respondí que nuestras respuestas eran complejas porque, en efecto, las respuestas son complejas. Uno de sus jefes había sido director general de una empresa de cómputo de la que yo poseía algunas acciones y señalé que los ahorros de mi retiro estarían en riesgo si ese director en particular hubiera sido lo bastante tonto para exigir respuestas simples cuando su equipo le presentara problemas de arquitectura de un chip o del diseño de un programa. La mujer no pareció convencida.

Quizá la evaluación parezca tan engañosamente simple porque para los que fuimos criados y educados en Estados Unidos las pruebas estandarizadas siempre han estado presentes y son inevitables. Nos aplicaron pruebas de logro en la primaria y la secundaria. La mayoría de nosotros presentó exámenes de admisión a la

educación media, algunos de manera repetida. Presentamos exámenes de lápiz y papel y pruebas computarizadas para obtener la licencia de conducir. Muchos presentamos exámenes para obtener la licencia que nos permitiría la entrada a un oficio o profesión. La evaluación se ha convertido en una parte rutinaria de nuestro vocabulario y discurso público. Durante muchos años la revista *Parade* publicó una columna de Marilyn vos Savant, quien, según la publicación, tiene el CI más elevado del país. En lugar de decir simple y llanamente que la señora Vos Savant es una persona sumamente inteligente, si en verdad lo es, los editores usan el vocabulario cotidiano del "CI" (un tipo de calificación de un tipo de prueba estandarizada). Los editores de *Parade* tienen una justificada confianza en que su referencia al CI deja claro lo que quieren decir, incluso si son muy pocos los lectores que tienen alguna idea del contenido de una prueba de CI o que están familiarizados con los argumentos sobre lo que en realidad miden dichas pruebas.

La etiqueta asignada a la señora Vos Savant señala otro problema: el poder retórico de las pruebas. Hacer referencia a su puntuación de CI en realidad no da más información para el lector común; es sólo otra forma de decir que es lista. Pero parece dar más peso a la aseveración, un matiz de credibilidad científica. (En este caso, el matiz es inmerecido; la afirmación de que cualquier persona es la más inteligente del país o tiene el mayor logro según la medición de alguna otra prueba, es absurda por todo tipo de razones, pero esa es otra cuestión.) Se supone que debemos creer que la inteligencia de dicha mujer no es sólo la opinión de los editores, sino algo que la ciencia ha validado.

Es verdad que la evaluación cuidadosa proporciona información de enorme valor acerca del logro de los estudiantes de la que de otro modo careceríamos (si no ¿para qué la haríamos?) y se basa en varias generaciones de investigación y desarrollo científico acumulado. Pero esa no es razón para hacer un uso acrítico de la

información de las pruebas. No es necesario ser un psicómetra para entender los principales problemas provocados por la evaluación del logro y para ser un usuario apto de la información que proporcionan las pruebas.

¿Cuáles son entonces algunas de las complicaciones que hacen que la evaluación y la interpretación de los resultados sean mucho menos sencillas de lo que cree la mayoría de la gente? Al principio pueden parecer desalentadoramente numerosas. Más o menos a las tres semanas del primer curso que impartí en Harvard, una alumna levantó la mano y me espetó: "¡Estoy verdaderamente frustrada!". Me quedé perplejo; el curso era nuevo y poco convencional y yo no tenía idea de qué tan bien funcionaría. Al principio esta alumna había sido una de las más entusiastas; si incluso ella se estaba desanimando, entonces yo veía problemas en el horizonte. Le pregunté a qué se debía su frustración, y respondió: "Cada día que venimos, unas cuantas respuestas sencillas se van por la alcantarilla. ¿Cuándo vamos a hacer algo para reemplazarlas?". Con un optimismo que en realidad no sentía, le dije que en el curso del semestre adquiriría una nueva comprensión de las pruebas que sería más compleja pero mucho más razonable y útil. Por fortuna, tanto ella como sus compañeros lo hicieron. Muchos otros estudiantes me han dicho que aprender sobre la evaluación se parece un poco a aprender un idioma extranjero, intimidante al principio pero cada vez más sencillo con la práctica.

De las muchas dificultades que conlleva la evaluación educativa, la fundamental, y la que en última instancia es la raíz de muchas ideas erróneas sobre las calificaciones, es el hecho de que los resultados por lo regular no proporcionan una medida directa y completa del logro educativo. Más bien, son medidas incompletas, sucedáneas de medidas más completas que usaríamos en condiciones ideales pero a las cuales por lo general no tenemos acceso. Existen dos razones para considerar a las pruebas de logro como

incompletas. Una, enfatizada durante más de medio siglo por los desarrolladores meticulosos de las pruebas estandarizadas, es que esos instrumentos sólo pueden medir un subconjunto de las metas de la educación. La segunda es que incluso en la evaluación de las metas que pueden medirse bien, las pruebas por lo general son muestras muy pequeñas de conducta que usamos para estimar el dominio que tienen los estudiantes de campos muy grandes de conocimiento y habilidad. Como se explica en los siguientes capítulos, una prueba de logro es en muchos sentidos como una encuesta política en la que se usan las opiniones de un pequeño número de votantes para predecir el voto posterior de muchísimas personas más. Esos hechos generan la mayor parte de las complejidades que se explican en este libro.

Al final haré varias sugerencias para usar razonable y fructíferamente las pruebas y sus resultados, muchos de los cuales se basan en última instancia en un solo principio: no tratar "la calificación obtenida en la prueba" como sinónimo de lo "que se aprendió". La calificación de una prueba es sólo un indicador de lo aprendido por un alumno, que si bien en muchos sentidos es de excepcional utilidad, es inevitablemente incompleto y algo propenso al error.

En ocasiones, una consecuencia inquietante de lo incompleto de las pruebas es que diferentes instrumentos arrojan a menudo resultados incongruentes. Por ejemplo, durante más de tres décadas el gobierno federal ha financiado una evaluación a gran escala de los alumnos de todo el país denominada Evaluación Nacional del Progreso Educativo (*National Assessment of Educational Progress*, NAEP), que en general se considera el mejor parámetro individual del logro de la juventud estadounidense. En realidad son dos las evaluaciones NAEP: una de ellas, la principal, diseñada para hacer un informe detallado en cualquier año dado, y otra diseñada para ofrecer estimaciones más sistemáticas de las tendencias de largo plazo. Las dos muestran que el logro en matemáticas

ha mejorado para cuarto y octavo grados, sobre todo para el cuarto grado, donde el incremento es uno de los cambios en el desempeño (hacia arriba o hacia abajo) más rápidos que se hayan registrado nunca en el país. Pero la tendencia ascendente de la NAEP principal ha sido notablemente más rápida que la mejora en la NAEP de la tendencia a largo plazo. ¿Por qué? Porque las pruebas hacen una medición algo diferente de las matemáticas, ya que toman muestras distintas de conducta del enorme dominio del logro en matemáticas y el progreso en el desempeño de los alumnos varía de un componente de las matemáticas a otro. Dichas discrepancias son algo frecuentes y no necesariamente indican que haya algo "malo" en cualquiera de las pruebas (aunque podría ser el caso); es posible que se deban simplemente a que las diferentes pruebas miden muestras algo distintas de conocimiento y habilidades. Pero esas discrepancias no son motivo para hacer a un lado las calificaciones obtenidas; más bien indican la necesidad de usarlas con cautela, por ejemplo, de considerar múltiples fuentes de información sobre el desempeño y prestar poca atención a las modestas diferencias en los resultados.

Sin embargo, hay casos en que una discrepancia entre dos pruebas puede señalar que hay algo mal y, en el mundo actual, un ejemplo de particular importancia puede presentarse en el caso de las pruebas de alto impacto, es decir, cuando se hace responsables a los educadores, a los estudiantes o a ambos de las calificaciones obtenidas. Cuando las calificaciones o resultados tienen consecuencias graves, las puntuaciones obtenidas en el examen que importa suelen aumentar mucho más rápido que en otras pruebas. Un buen ejemplo se encuentra en la experiencia en Texas durante la gestión de George Bush como gobernador. En ese tiempo, el estado utilizaba la Evaluación Texana de las Habilidades Académicas (*Texas Assessment of Academic Skills*, TAAS) para evaluar a las escuelas y se exigía a los estudiantes de preparatoria que la aprobaran

para poder recibir el diploma. Los estudiantes texanos mostraron un progreso espectacularmente mayor en esa prueba que en la Evaluación Nacional del Progreso Educativo.[2] El problema no es exclusivo de Texas, se han encontrado discrepancias similares en otros estados y localidades. En esos casos, la pregunta obvia (que abordamos en cierto detalle más adelante en el libro) es si se inflaron las calificaciones en la prueba de alto impacto y si por ende son un indicador engañoso de lo que en realidad aprendieron los estudiantes. Esta pregunta es lógicamente inevitable, pero en la práctica es común que sea ignorada incluso por personas que deberían advertir que es erróneo hacerlo por el simple hecho de que el problema de la inflación de las calificaciones en el mejor de los casos es una molestia y en el peor una amenaza. (Lo último es una de las razones que explican la escasez de estudios sobre este problema. Imagínese a un investigador que se acerca a un secretario de educación o al director estatal de evaluación para decirle: "Me gustaría tener acceso a los datos de su prueba para investigar si los incrementos en los resultados que ha estado reportando a la prensa y al público están inflados". La propuesta no es atractiva.)

Incluso una sola prueba puede arrojar resultados variables. Así como las encuestas tienen un margen de error, lo mismo sucede con las pruebas de logro. Los estudiantes que presentan más de una versión de la prueba por lo general obtienen calificaciones diferentes.

Por ejemplo, cuando presentan más de una vez la prueba SAT de admisión a la universidad, es común encontrar fluctuaciones de 20 a 30 puntos en sus calificaciones en los componentes verbal y matemático de una ocasión a otra. Esas fluctuaciones surgen porque si bien las formas del examen están diseñadas para ser equivalentes, su contenido es distinto y el sustentante puede estar de suerte una ocasión o la siguiente. Las fluctuaciones también ocurren porque los estudiantes tienen días buenos y malos: tal vez el

sustentante estaba demasiado nervioso para poder dormir bien la noche previa al primer examen o tenía dolor de estómago en el segundo. De modo que no tiene sentido dar tanta importancia a las pequeñas diferencias en las calificaciones y el Consejo de Universidades, auspiciador del SAT, exhorta a los usuarios a que no lo hagan. Las pequeñas diferencias en las calificaciones simplemente no son lo bastante sólidas para ser dignas de confianza. Algunos programas estatales y locales de evaluación ofrecen información acerca de este margen de error a los padres, para ello les entregan un reporte (que sospecho que muchos encuentran demasiado intrincado) que demuestra que el nivel de logro de su hijo en realidad cae en algún punto en un rango alrededor del resultado que recibió.

Luego surge el problema de saber cómo informar sobre el desempeño en una prueba. La mayoría de nosotros creció en un sistema escolar con algunas reglas simples, aunque arbitrarias, para calificar los exámenes, tales como "con 90 por ciento de aciertos obtienes una A". Pero si se sustituyen algunas preguntas difíciles por otras más sencillas, o al revés (y este tipo de variaciones ocurre aunque la gente trate de evitarlo), y "90 por ciento de aciertos" ya no significa el nivel de dominio que significaba antes. En cualquier caso ¿qué es una "A"? Sabemos que para obtener una calificación de "A" en una clase se requiere mucho más que en otra. Por consiguiente, los psicómetras han tenido que crear escalas para informar sobre el desempeño en las pruebas. Esas escalas son de muchos tipos distintos. La mayoría de los lectores habrá encontrado escalas numéricas arbitrarias (por ejemplo, la escala SAT que va de 200 a 800); escalas referidas a normas, que comparan a un estudiante con una distribución de estudiantes, quizá una distribución nacional (por ejemplo, equivalentes de grado y rangos percentiles); y los estándares de desempeño que dominan en el momento, los cuales descomponen toda la distribución de desempeño en unos cuantos intervalos basados en juicios de lo que los estudiantes

deben ser capaces de hacer. Esas diversas escalas tienen diferentes relaciones con el desempeño crudo en la prueba, por lo que a menudo proporcionan visiones distintas del desempeño.

Además, a veces una prueba no funciona como debería. Un instrumento puede estar sesgado y producir de manera sistemática estimaciones incorrectas del desempeño de un determinado grupo de alumnos. Por ejemplo, una prueba de matemáticas que exija la lectura de un texto difícil y la redacción de respuestas largas puede estar sesgada en contra de los estudiantes inmigrantes que son competentes en matemáticas pero que no han adquirido fluidez en el inglés. Esos casos de sesgo deben distinguirse de las simples diferencias en el desempeño que representan con exactitud el logro. Por ejemplo, si los estudiantes pobres de una determinada ciudad asisten a malas escuelas, es probable que una prueba sin sesgo alguno les otorgue menores calificaciones debido a que la enseñanza inferior que recibieron impidió su aprendizaje.

Esas complicaciones no siempre tienen una solución sencilla y clara. Hace años, un destacado y muy serio reformador de la educación asistió a una reunión del Comité de Evaluación y Examinación del Consejo Nacional de Investigación, el mismo grupo visitado por la contrariada representante de un grupo de empresas que mencioné antes. Solicitó que el consejo financiara un panel de estudio que generara el diseño óptimo para un programa de evaluación. Para su evidente molestia, le respondí que eso no sería un proyecto acertado porque no existe un diseño óptimo. Más bien, el diseño de un programa de evaluación es un ejercicio de negociaciones y compromisos, y el juicio acerca de qué compromiso es mejor dependerá de detalles concretos, como los usos particulares que se harán de los resultados de las evaluaciones. Por ejemplo, los diseños de evaluación que son mejores para proporcionar información descriptiva acerca del desempeño de grupos (como escuelas, distritos, estados e incluso naciones enteras) no son apropiados

para sistemas en que debe compararse el desempeño de estudiantes individuales. Añadir tareas largas, complejas y exigentes a una evaluación tal vez amplíe el rango de habilidades que pueden evaluarse, pero al costo de hacer menos fidedigna la información acerca de estudiantes en particular. Cualquier diseño que ofrece ganancias en un frente seguramente impone costos en otro, y la tarea de hacer un diseño razonable implica ponderar con cuidado los intercambios inevitables.

Pero son pocas las personas que participan en la elaboración o la elección de pruebas. La mayoría de quienes están interesados en la evaluación (padres, políticos, ciudadanos preocupados por la calidad de las escuelas o la preparación de la fuerza laboral) sólo quieren hacer buen uso de sus resultados. ¿En qué medida es en verdad necesario que confronten este maremágnum de complicaciones?

La evaluación es por naturaleza una empresa altamente técnica que descansa en una base matemática compleja, buena parte de la cual no es entendida ni siquiera por los científicos cuantitativos sociales de otros campos. Muchos de los reportes técnicos colocados en internet por los departamentos estatales de educación y por otras organizaciones que patrocinan las pruebas enfrentan a los lectores con términos técnicos apabullantes y, en algunos casos, con las matemáticas aún más intimidantes que los formalizan. Esto crea el desafortunado malentendido de que los principios de la evaluación están fuera del alcance de la mayoría de la gente. Las matemáticas son esenciales para el diseño y la operación apropiada de los programas de evaluación, pero uno no necesita entender los principios que subyacen a los usos acertados de las pruebas y a las interpretaciones razonables de sus resultados.

Pero los principios y conceptos centrales sí son esenciales. Sin una comprensión de lo que es validez, confiabilidad, sesgo, escalamiento y establecimiento de estándares, por ejemplo, uno no puede entender por completo la información arrojada por las

pruebas o encontrar soluciones inteligentes a las amargas controversias actuales sobre la evaluación educativa. Muchas personas simplemente descartan esas complejidades y les restan importancia justamente porque parecen técnicas y esotéricas. Sospecho que esto fue en parte el problema con la representante del grupo de empresas que mencioné antes. Esta propensión a asociar lo críptico con lo no importante es absurdo y pernicioso. Cuando estamos enfermos, casi todos confiamos en tratamientos médicos que reflejan un conocimiento complejo y esotérico de toda suerte de procesos fisiológicos y bioquímicos que pocos comprendemos. Sin embargo, pocos de nosotros le diríamos a nuestro médico que es imposible que su conocimiento (o el de los investigadores biomédicos que diseñaron los medicamentos que estamos tomando) sea importante porque resulta ininteligible para nuestros oídos profanos. Tampoco descartamos la insondable ingeniería utilizada en los sistemas de control de las modernas aeronaves o, para el caso, en las computadoras que controlan nuestros automóviles. Ignorar las complejidades de la evaluación educativa conduce a la gente a errores importantes acerca del desempeño de los estudiantes, las escuelas y los sistemas escolares. Y aunque las consecuencias de esos malos entendidos quizá no parezcan tan funestas como los aviones que caen del cielo, son lo bastante graves para los niños, para sus maestros y para la nación, cuyo bienestar depende de una ciudadanía bien educada. ■

Capítulo 2
¿Qué es una prueba?

El 10 de septiembre de 2004, una encuesta aplicada por Zogby International a 1,018 probables votantes mostró que George W. Bush aventajaba a John Kerry con cuatro puntos porcentuales en la campaña para la elección presidencial. Esos resultados constituyeron una predicción razonablemente buena: dos meses más tarde, Bush ganó con un margen aproximado de 2.5 por ciento.

Este tipo de encuestas es tan común que son pocos los que piensan en otra cosa que nos sean sus resultados finales. A veces, las encuestas cometen errores monumentales (el ejemplo clásico es el triunfo de Truman sobre Dewey en 1948, pero otro más reciente y más espectacular fue la inesperada victoria de Hamas en las elecciones palestinas de 2005), y eso suele dar lugar a discusiones sobre la manera en que se realizan las encuestas y a qué puede deberse que las encuestas en cuestión se alejaran tanto del blanco. Sin embargo, es más habitual que el lector promedio de los diarios no preste atención a los detalles técnicos de dichos estudios.

No obstante, los principios en que se basan las encuestas no sólo son fundamentales para las ciencias sociales sino que además proporcionan una forma práctica de explicar la operación de las pruebas de logro.

¿Habría alguna razón para que le interesara cómo votaron esos 1,018 participantes en la encuesta de Zogby? Son raros los casos (como el de Florida en 2000) en que uno podría preocuparse por los votos de un número tan reducido de personas. Sin embargo,

casi siempre las personas muestreadas para una encuesta son demasiado pocas para influir en el resultado de la elección real. El número total de votos emitidos en la elección de 2004 excedió los 121 millones y el margen de Bush fue mayor a tres millones de votos. Si los 1,018 probables votantes encuestados por Zogby hubiesen votado por uno solo de los dos candidatos, el cambio en el conteo final (alrededor de 500 votos) habría sido demasiado pequeño para advertirlo.

Así es que ¿por qué deberíamos preocuparnos por esos 1,018 individuos? Porque en conjunto representan a los 121 millones de personas que *sí* nos interesan, y las 1,018 personas nos permiten predecir (en este caso bastante bien) la conducta del grupo mayor. No podemos medir directamente la intención de voto de alrededor de 121 millones de personas porque hacerlo resultaría prohibitivamente costoso y tardado. Además, no nos interesa tanto la intención de voto de esos 121 millones de individuos; nos interesa saber cómo votarán en realidad y no hay dinero o esfuerzo que nos lo diga con certeza antes del día de las elecciones. Enfrentados con la imposibilidad de medir directamente lo que en realidad nos importa, confiamos en una medida sustituta más práctica: encuestamos a un pequeño número de personas acerca de sus intenciones y usamos sus respuestas para predecir la información que no podemos obtener y que es en realidad la que queremos acerca de toda la población de votantes.

Nuestra habilidad para hacer esta predicción a partir de los resultados de la encuesta depende de muchas cosas. Depende del diseño de la muestra, que debe elegirse con cuidado para que represente a la población mayor de probables votantes. Si Zogby sólo hubiese muestreado a individuos de Utah o, por el contrario, de Massachusetts, la muestra no habría sido una buena representación de los votantes del país y habría dado lugar a una predicción errónea. (Cuando yo vivía en la zona metropolitana de Washington,

teníamos nuestra propia variante del dicho "Lo que pase en Ohio, pasa en el resto de la nación". El nuestro decía "Lo que pase en el Distrito de Columbia, pasa en Massachusetts".) Los encuestadores no cometerían un error tan evidente, pero errores más sutiles en el diseño de la muestra, algunos de los cuales son imprevistos, pueden sesgar gravemente los resultados. La precisión también depende de la manera en que se formulan las preguntas de la encuesta; existe mucha evidencia de que incluso cambios que parecen menores en la redacción de las preguntas pueden tener efectos considerables en las respuestas de los encuestados. Por ejemplo, un estudio comparó las respuestas al siguiente par de preguntas:

> *Pregunta original:* "¿Cuál es el número promedio de días de la semana que consume mantequilla?"
> *Pregunta revisada:* "La siguiente pregunta es sólo acerca de la mantequilla. Sin incluir a la margarina, ¿cuál es el número promedio de días de la semana que consume *mantequilla*?"

Las preguntas se plantearon a grupos equivalentes de participantes. Uno podría pensar que la aclaración en la segunda pregunta era innecesaria, pero las respuestas indicaron otra cosa. De los encuestados a los que se planteó la pregunta original, 33 por ciento respondió que cero días y 23 por ciento dijo que siete días. De la muestra equivalente a la que se formuló la pregunta revisada, 55 por ciento respondió que cero días y sólo 9 por ciento dijo que siete días.[1]

Por último, la exactitud depende de la capacidad o la disposición de los encuestados a proporcionar la información solicitada. Los individuos de la muestra tal vez estén dispuestos a responder pero carecen de la información que se les pide (como cuando se pregunta a los estudiantes acerca del ingreso de los padres). Quizá se nieguen a responder, como yo suelo hacerlo

cuando las empresas de investigación de mercados me llaman a la hora de la comida. Puede ser que respondan pero que proporcionen información inexacta. A los investigadores les preocupa lo que llaman el "sesgo de deseabilidad social", la tendencia de algunos encuestados a proporcionar respuestas que son socialmente aceptables pero inexactas. Podría esperarse que esto sucediera si se les pregunta acerca de conductas o actitudes indeseables (como el sesgo racial), pero también se encuentra un reporte excesivo de conductas, actitudes y estatus socialmente deseables. Durante más de medio siglo se ha documentado la gravedad de este sesgo. Por ejemplo, un estudio publicado en 1950 documentó un reporte considerablemente excesivo de diversos tipos de conducta socialmente deseable. Treinta y cuatro por ciento de los encuestados dijeron que habían contribuido a una determinada organización local de beneficencia cuando no era así, y entre 13 y 28 por ciento de los encuestados dijeron que habían votado en varias elecciones en que no lo habían hecho.[2] Pero cuando todo sale bien, los resultados de una minúscula muestra proporcionan una estimación razonable de los hallazgos que uno habría obtenido de la población como un todo. Ese fue el caso de la encuesta de Zogby.

Las pruebas de logro educativo son análogas a esta encuesta de Zogby en el sentido de que son un estimado de una medida mejor y más completa que no podemos obtener. En la mayoría de los casos, el consumidor de las calificaciones obtenidas en una prueba (un padre que quiere conocer el desempeño de su hijo, un administrador que busca las áreas fuertes y débiles en el desempeño de las escuelas, un político que quiere criticar a las escuelas o disfrutar el brillo de su progreso) desea sacar conclusiones acerca del manejo que tienen los estudiantes de una gran variedad de conocimientos y habilidades. En el caso de una prueba de fin de cursos, esto podría ser algo como el manejo de los conceptos y habilidades en álgebra básica. En otros casos, el campo de conocimiento podría

ser incluso más amplio. Por ejemplo, muchos estados aplican pruebas de matemáticas que se diseñaron para proporcionar información acerca de la destreza acumulativa en las matemáticas a lo largo de muchos grados.

Por lo general, quienes están en el oficio llaman *dominio* a toda la variedad de habilidades y conocimientos acerca de los cuales la prueba hace una estimación –lo que es análogo a la estimación que hizo la encuesta de Zogby de los votos de toda la población. Así como no es factible que los encuestadores obtengan información de toda la población, tampoco es posible que una prueba mida de manera exhaustiva todo un dominio porque estos suelen ser demasiado grandes. En lugar de ello elaboramos una prueba de logro que cubra una pequeña muestra del dominio, justo como el encuestador selecciona una pequeña muestra de la población. Y como en el caso de la encuesta, el tamaño pequeño no es la única limitación de la muestra que medimos. De la misma forma en que un encuestador no puede medir directamente una futura conducta de votación, hay algunos aspectos de las metas de la educación que las pruebas de logro no pueden medir.

La analogía entre la encuesta política de Zogby y una prueba de logro falla en un aspecto: esta encuesta particular se utilizó para predecir algo que se situaba en el futuro y que por ende era imposible conocer, mientras que las pruebas de logro por lo general (pero no siempre) se utilizan para medir algo que los estudiantes ya saben. Sin embargo, esta diferencia es más aparente que real. En ambos casos se utiliza una pequeña muestra para calcular un conjunto mucho mayor: en un caso, la conducta de un grupo mayor de personas y, en el otro, un conjunto mayor de conocimiento y habilidades. Y en ambos casos (aunque por razones algo diferentes) el conjunto mayor no puede medirse de manera directa y exhaustiva.

Los resultados de una prueba de logro (la conducta de los estudiantes al responder a una pequeña muestra de preguntas) se

utilizan para estimar cómo sería el desempeño de los estudiantes en todo el dominio si pudiésemos medirlo directamente. En el caso de la encuesta de Zogby, nos interesa que los resultados nos permitan hacer un cálculo preciso de los votos posteriores de toda la población, pero en realidad no nos interesan los votos finales emitidos por los pocos integrantes de la muestra. La evaluación del logro es parecida. No nos preocupa demasiado el desempeño de los estudiantes en un reactivo específico de la prueba (se les llama *reactivos* porque no tienen que redactarse como preguntas), de la misma manera que no nos preocuparía el voto posterior de un único participante en la encuesta de Zogby. La importancia del reactivo, igual que la importancia del encuestado, radica en el conjunto mayor de conocimiento y habilidades que representa.

La precisión de las estimaciones que se basan en una prueba depende de varios factores. Así como la precisión de una encuesta depende del muestreo cuidadoso de los individuos, la precisión de una prueba depende del muestreo cuidadoso del contenido y las habilidades. Por ejemplo, si queremos medir la competencia matemática de los alumnos de octavo grado, debemos especificar a qué conocimiento y habilidades nos referimos por "matemáticas de octavo grado". Podríamos decidir que esto incluye habilidades en aritmética, medición, geometría plana, álgebra básica y análisis de datos y estadística, pero luego tendríamos que decidir *qué aspectos* del álgebra y de la geometría plana importan y cuánto peso debe darse a cada componente. ¿Es necesario que los estudiantes conozcan la fórmula cuadrática? Al final, terminamos con un mapa detallado de lo que las pruebas debieran incluir, y que a menudo se conoce como "especificaciones de la prueba" o "programa detallado de la prueba" y con una serie de reactivos como muestra de esas especificaciones escritas por el desarrollador.

Pero ese es apenas el principio. De la misma manera que la precisión de una encuesta depende a menudo de detalles al parecer

misteriosos sobre el planteamiento de las preguntas, la precisión de los resultados de la prueba depende de una multitud de detalles a menudo enigmáticos sobre la redacción de los reactivos, la redacción de los "distractores" (las respuestas erróneas en los reactivos de opción múltiple), la dificultad de los reactivos, la rúbrica (reglas y criterios) usada para calificar el trabajo de los alumnos, etcétera. Y así como la precisión de una encuesta depende de la disposición de los participantes a responder con franqueza, la precisión del puntaje de una prueba depende de la actitud de los examinados (por ejemplo, su motivación para tener un buen desempeño). También depende, como veremos más adelante, de la conducta de otros, en particular de los maestros. Si se presentan problemas en cualquiera de esos aspectos de la evaluación, los resultados de la pequeña muestra de conducta que constituye la prueba arrojarán estimaciones engañosas del manejo que tienen los estudiantes del dominio mayor. Nos alejaremos creyendo que, después de todo, Dewey vencerá a Truman o, para ser precisos, creeremos que de hecho ya lo hizo.

Esto podría llamarse el *principio de muestreo* de la evaluación: las calificaciones reflejan una pequeña muestra de conducta y sólo son valiosas en la medida en que apoyan las conclusiones acerca de dominios mayores de interés.

Este es tal vez el principio fundamental de las pruebas de logro. La incapacidad de entenderlo está en la raíz de muchas ideas erróneas generalizadas sobre los resultados de las pruebas. A menudo ha desviado los esfuerzos de los políticos por diseñar sistemas productivos de evaluación y rendición de cuentas. También ha desembocado en innumerables casos en que los maestros y otras personas llevan a cabo ejercicios malos de preparación para la prueba al enfocar la instrucción en la pequeña muestra que es evaluada, en lugar de hacerlo en el conjunto más amplio de habilidades cuyo manejo se supone que debe señalar la prueba. Muchos

otros principios clave de la evaluación, así como algunos de los debates actuales más acalorados (en particular, el debate acerca de hacer responsables a los maestros de los resultados de las pruebas), son un resultado directo de esta realidad sobre el muestreo.

La construcción de una prueba hipotética ayudará a concretar este y otros principios esenciales de la evaluación. Suponga que usted es el editor de una revista y ha decidido contratar a algunos estudiantes universitarios como internos para que le ayuden. Recibe un gran número de solicitudes y determina que un criterio para elegir entre ellos es la riqueza de su vocabulario.

¿Cómo va a establecer qué aspirantes tienen vocabularios particularmente ricos? Si los conociera bien quizá tendría algún conocimiento de su vocabulario a partir de la experiencia acumulada en muchas discusiones en diversos contextos. Pero si no los conoce, tiene poco en qué basarse. Podría someter a cada uno a una breve entrevista, pero probablemente eso arrojaría información escasa y desigual de un aspirante a otro. Su conversación con los candidatos podría fluir en direcciones muy diferentes, lo que brindaría a algunos de ellos mayor oportunidad que a otros de demostrar un vocabulario rico. Seguramente no querría dejar escapar a una buena aspirante porque resultó que su conversación con ella terminó por enfocarse en los Medias Rojas más que en el equilibrio del mercado.

Una opción clara es aplicar a los candidatos una prueba de vocabulario para complementar lo que sabe de ellos a partir de otras fuentes como la entrevista. Esto tiene varias ventajas obvias. La prueba, a diferencia de las entrevistas, estaría diseñada específicamente para obtener información sobre el vocabulario de cada aspirante. También sería constante de un aspirante a otro. Cada aspirante enfrentaría las mismas tareas y, por lo tanto, su desempeño no estaría sujeto a los caprichos de la conversación. La candidata que terminó hablando sobre los Medias Rojas presentaría la misma prueba que quien discutió el déficit del mercado.

Esta es la razón de la *estandarización*. La gente hace un uso incorrecto del término *prueba estandarizada* (a menudo con menosprecio) para referirse a todo tipo de cosas: pruebas de opción múltiple, pruebas diseñadas por empresas comerciales, etcétera. De hecho, sólo quiere decir que la prueba es idéntica. En concreto, sólo significa que todos los examinados enfrentan las mismas tareas, que se aplican y se califican de la misma manera. La motivación para la estandarización es sencilla: evitar factores irrelevantes que podrían distorsionar las comparaciones entre individuos. Si usted presentara a sus aspirantes tareas no estandarizadas, podría concluir erróneamente que los que tuvieron que definir las palabras más sencillas (o cuyas pruebas fueron calificadas con criterios más indulgentes) tenían un vocabulario más rico. La estandarización también tiene desventajas, en particular para algunos estudiantes con discapacidades o cuya competencia en el idioma del examen es limitada, pero en general, es más probable que las evaluaciones estandarizadas proporcionen información comparable que las no estandarizadas.

Supongamos entonces que decidió aplicar una prueba estandarizada de vocabulario y que, por ende, necesita elaborarla. Tendría que enfrentar entonces una grave dificultad: aunque a muchos padres les parezca sorprendente a la luz de su propia experiencia, el adolescente típico tiene un vocabulario enorme. Una reciente estimación muy respetada es que el graduado común de bachillerato tiene un vocabulario funcional de alrededor de 11 mil palabras mientras que el universitario representativo tiene cerca de 17 mil.[3] Es claro que no va a sentar a los aspirantes y preguntarles acerca de 11 mil o 17 mil palabras. Ni siquiera es práctico preguntarles acerca del subconjunto de los muchos miles de palabras que pueden necesitar en realidad en el trabajo en su revista. Ese subconjunto también es demasiado grande.

Lo que tendría que hacer es seleccionar una muestra de esas miles de palabras para usarlas en su prueba. En la práctica, puede

hacer una estimación razonablemente buena de la riqueza relativa del vocabulario de los candidatos evaluándolos con una pequeña muestra de palabras siempre que se elijan con cuidado. Supongamos que en este caso usted utilizará cuarenta palabras, que no sería un número inusual en una verdadera prueba de vocabulario.

La selección de las palabras que incluirá en la prueba se convierte entonces en la primera clave para obtener información útil de la prueba. La figura 2.1 presenta las tres primeras palabras de cada una de tres listas entre las cuales podría seleccionar al elaborar su prueba. Las palabras adicionales en cada una de las tres listas que no aparecen en la figura 2.1 son similares a las presentadas en términos de dificultad y frecuencia de uso. ¿Qué lista utilizaría? Seguro no sería la lista A, ya que contiene palabras especializadas que casi nunca se usan y que muy pocos de sus aspirantes conocerían. (La verdad sea dicha, para elaborar la lista A revisé la versión completa de un diccionario para encontrar palabras que no recordaba haber visto antes. Para los curiosos, *silícula* es el término botánico que se refiere a las plantas que tienen cápsulas de semillas biválvicas, como las plantas de mostaza; *vilipendiar* es un término arcaico que significa menospreciar, y *epimisio* es la membrana exterior que recubre al músculo.) Dado que prácticamente ninguno de sus solicitantes conocería las palabras de la lista A, la prueba sería demasiado difícil para ellos. Todos recibirían una calificación de cero o cercana a cero y eso haría que la prueba fuese inútil: usted no obtendría información útil sobre la riqueza relativa de su vocabulario.

Una vez más, la lista B no es mejor. Es muy probable que todos sus postulantes conozcan las definiciones de *baño*, *viaje* y *alfombra*. Cualquiera podría obtener una calificación perfecta o casi perfecta. En este caso no se aprendería nada, porque la prueba sería demasiado sencilla.

Figura 2.1. Tres palabras de cada una de tres listas hipotéticas

A	B	C
silícula	baño	indolente
vilipendiar	viaje	menospreciar
epimisio	alfombra	minúsculo

Por lo tanto, elaboraría su prueba a partir de la lista C, la cual incluye palabras que algunos aspirantes conocen y otros no. Podría decir que algunas de esas palabras no son de la dificultad correcta (tal vez *minúsculo* sea demasiado fácil para los universitarios), pero esa es una cuestión empírica que un autor de pruebas cuidadoso respondería sometiendo los posibles reactivos a una prueba piloto. Usted desea terminar con una lista de palabras que algunos de los aspirantes, pero no todos, puedan definir correctamente. Existen razones técnicas para elegir el rango específico de dificultad de los reactivos, pero para los propósitos actuales es suficiente ver que desea reactivos de dificultad moderada que algunos de los estudiantes responderán correctamente y otros de manera errónea.

En este ejemplo, es claro el principio de muestreo de la evaluación. Usted está interesado en el manejo que tienen los aspirantes de un gran número de palabras (el dominio), pero la evidencia con que cuenta es su conocimiento de la pequeña muestra incluida en la prueba. En este caso, la muestra está conformada por las 40 palabras de la prueba y el dominio que representa esa muestra son los vocabularios de los aspirantes, que incluyen miles de palabras. Usted habría evaluado tal vez una de cada 300 o 400 palabras que conocen los aspirantes. Es evidente por qué el desempeño en los reactivos individuales incluidos en esta prueba no debería ser el

foco de interés: las 40 palabras concretas que se evaluaron no importan mucho porque son apenas una gota en la cubeta. Lo que importa es la estimación que proporcionan del conocimiento del gran dominio del que fueron muestreadas.

Pero ¿es el muestreo siempre un problema tan grave como en este ejemplo inventado? Hay casos en que no es así. De hecho, existen casos raros en que es posible evaluar todo un dominio sin ningún muestreo. Por ejemplo, he tenido alumnos interesados en evaluar las incipientes habilidades de alfabetización de jóvenes estudiantes. En los lenguajes alfabéticos, una de dichas habilidades es el reconocimiento de las letras, y puesto que las letras por aprender no son muchas, resulta sencillo examinar el conocimiento que tienen los colegiales de todas ellas. En este caso, el contenido evaluado es el dominio, no una muestra del mismo.

Sin embargo, en su mayor parte las pruebas que resultan de interés para los políticos, la prensa y el público en general implican un muestreo extenso porque están diseñadas para medir dominios de proporciones considerables, que van del conocimiento adquirido a lo largo de un año de estudio de un tema a la pericia acumulada en el material estudiado a lo largo de varios años. Las pruebas del Sistema de Evaluación Integral de Massachusetts (*Massachusetts Comprehensive Assessment System*, MCAS) pueden servir como ejemplo. Los estudiantes de Massachusetts deben obtener una calificación aprobatoria en las pruebas de décimo grado de matemáticas y artes del lenguaje para recibir el diploma de bachillerato. La prueba de matemáticas de décimo grado abarca cinco áreas de conocimiento de la materia estudiadas a lo largo de varios años de escolaridad: sentido de los números y operaciones (que incluye aritmética); números, relaciones y álgebra; geometría; medición; y análisis de datos, estadística y probabilidad. En la primavera de 2005, las calificaciones de los estudiantes en matemáticas se basaron en 42 reactivos, un promedio de menos de nueve reactivos

por cada una de las cinco áreas de matemáticas. (Otros reactivos contribuyeron a las calificaciones para las escuelas, pero no las de los estudiantes.)[4]

Obviamente, esto es un grado austero de muestreo, pero si se cumplen ciertas condiciones, 42 reactivos constituyen una muestra lo bastante grande para proporcionar una buena cantidad de información útil. Un requisito es obvio: los reactivos deben ser elegidos de manera concienzuda para representar el contenido apropiado, justo como los 1,018 participantes en la encuesta de Zogby tuvieron que seleccionarse meticulosamente para representar a los probables votantes. Por ejemplo, si no se incluyó la trigonometría en el programa de matemáticas de décimo grado, la muestra evaluada no debe incluir reactivos de ese tema. Ya se mencionaron otros dos elementos que hay que considerar: la importancia de la estandarización y un nivel apropiado de dificultad.

La cuestión de la dificultad requiere de cierta discusión porque está en la raíz de varios errores graves en los debates actuales sobre la evaluación. Para su prueba de vocabulario usted eligió palabras con un nivel moderado de dificultad porque necesitaba reactivos que *discriminaran* entre los estudiantes con vocabularios amplios y los que tienen vocabularios reducidos. En este contexto, el término *discriminar* no tiene connotaciones negativas ni implica injusticia para un individuo o un grupo. Los reactivos y las pruebas que discriminan son sencillamente las que distinguen entre los estudiantes con mayor conocimiento y habilidades en lo que uno quiere medir y los que tienen menos. En este caso, usted quiere reactivos con mayor probabilidad de ser respondidos por los alumnos con vocabularios más ricos. Los reactivos que son demasiado difíciles o demasiado sencillos no pueden discriminar (prácticamente ningún aspirante conocerá el significado de *vilipendiar*), pero los reactivos de dificultad moderada también pueden fallar en la discriminación si miden algo más que la competencia

para cuya medición se diseñó la prueba. Sin reactivos que discriminen usted no tendría base para utilizar el desempeño en la prueba para clasificar a los aspirantes en términos de la estimación de sus vocabularios.

En el debate público es común que se arremeta contra los psicómetras por buscar reactivos que discriminen. En ocasiones se escuchan afirmaciones de que el uso de tales reactivos "crea ganadores y perdedores" y que eso, y no la medición exacta, es la meta de quienes diseñan ciertos tipos de prueba. Sin embargo, no hay nada nocivo en la elección de reactivos que discriminen, estos son necesarios si se desea hacer inferencias acerca de la competencia *relativa*. Eso quedó claro en el ejemplo del vocabulario: usted eligió reactivos que discriminaban para poder calcular el vocabulario relativo de los aspirantes. Usted no *creó* las diferencias de vocabulario entre los candidatos al hacer esa elección; simplemente hizo posible que la prueba *revelara* las diferencias ya existentes.

Existen otros usos para los cuales son adecuados los reactivos que no discriminan. Por ejemplo, es posible que una maestra quiera saber si su grupo aprendió la ortografía de una lista de palabras que le enseñó la semana pasada, en este caso se sentiría feliz si los reactivos de su examen no discriminaran en absoluto, es decir, si casi todos los estudiantes los respondieran bien. La clave es la inferencia particular que la maestra quiere hacer a partir de las calificaciones obtenidas en la prueba. Si hubiese usado reactivos no discriminantes no tendría base para hacer una inferencia sobre la competencia relativa, pero sí la tendría para hacer una inferencia sobre el manejo de ese material en concreto.

En la práctica, decidir cuándo se necesitan reactivos que discriminen es un poco más peliagudo de lo que puede parecer a primera vista, y en el mundo de la política educativa se ha generalizado un malentendido sobre este punto. Muchos programas actuales de evaluación se diseñaron en parte para determinar si los estudiantes

alcanzaron un conjunto de estándares de desempeño, como el estándar de "competente" exigido por la ley «Que ningún niño se quede atrás» (*No Child Left Behind*, NCLB). Muchos políticos y educadores sostienen –erróneamente– que esto es análogo a una prueba semanal de ortografía ya que sólo les interesa saber si los estudiantes dominaron lo que se necesita para alcanzar el nivel competente. Si no quieren diferenciar entre los niños más allá de distinguir entre los que son o no son competentes, ¿por qué necesitarían reactivos que discriminen? Pero incluso si sólo se está interesado en la distinción binaria entre competente y no competente –lo cual, en mi experiencia, pasa con muy poca gente– la complicación estriba en que "competente" es sólo un punto arbitrario en un continuo de desempeño; no indica la maestría de la totalidad de un conjunto diferenciado de habilidades. Para obtener información confiable acerca de qué niños alcanzaron en realidad el estatus de competente, se necesitan reactivos de examen que discriminen bien entre los niños cuyo manejo se acerca a ese nivel de competencia. (Un problema todavía mayor es decidir dónde poner el punto de corte que separe los fracasos de los éxitos "competentes". Esto se analiza en el capítulo 8.)

Volviendo a la prueba de vocabulario, ¿qué habría sucedido si usted hubiera elegido las palabras de manera diferente a la vez que las mantenía en el mismo nivel de dificultad y discriminación? La figura 2.1 sólo muestra las tres primeras palabras de cada una de las tres listas. Sin embargo, esas listas podrían contener cientos de palabras de dificultad y frecuencia de uso comparables. Usted podría elegir un conjunto de 40 de la lista C y yo podría elegir otras 40 de la misma lista. ¿Deberíamos preocuparnos?

Para concretar, suponga que eligió las tres palabras mostradas en la figura, por lo que su prueba incluyó *mezquino*, *menospreciar* y *minúsculo*. Resulta que yo también elegí las dos últimas, pero en lugar de *mezquino* elegí *indolente*. Esto se muestra en la figura 2.2. En

Figura 2.2. Sustitución de palabras en una lista hipotética de palabras

A	B	C
silícula	baño	~~indolente~~ mezquino
vilipendiar	viaje	menospreciar
epimisio	alfombra	minúsculo

aras del análisis, suponga que esas dos palabras son de igual dificultad. Es decir, si aplicamos a un gran número de estudiantes reactivos sobre ambas palabras, la misma proporción los respondería correctamente.

¿Cuál sería el impacto de aplicar mi prueba en lugar de la suya? Con un número de aspirantes lo bastante grande el resultado promedio no se vería afectado en absoluto, porque la dificultad de las dos palabras en cuestión es la misma. Sin embargo, los resultados de algunos estudiantes individuales *sí* se verían afectados. Incluso entre los alumnos con vocabularios comparables, algunos pueden conocer *indolente* pero no m*ezquino y* a la inversa.

Esto es un ejemplo del *error de medición*, el cual se refiere a la inconsistencia entre los resultados de una medición con la siguiente. Hasta cierto punto, la clasificación de sus aspirantes dependerá de qué palabras eligió de la columna C; y si los examinara de manera repetida usando diferentes versiones de la prueba, las clasificaciones variarían un poco. Casi cualquiera que haya presentado pruebas de admisión a la universidad, o que tenga hijos o alumnos que lo hayan hecho, está familiarizado con esto. Muchos estudiantes presentan más de una vez exámenes de admisión a la universidad como el SAT o el ACT, y sus calificaciones casi siempre varían un poco, aun cuando las pruebas se construyeron de manera concienzuda para que las formas fueran comparables. Una fuente de esta

inestabilidad en los resultados es que los autores de las pruebas eligieron diferentes reactivos para cada versión, y una versión puede resultar un tanto más ventajosa o desventajosa que la siguiente para un determinado alumno, incluso si la dificultad promedio de las distintas formas es la misma para todos los alumnos examinados. Otra fuente de inestabilidad es la fluctuación que puede ocurrir a lo largo del tiempo, incluso si los reactivos son los mismos. Los estudiantes tienen días buenos y malos. Por ejemplo, un estudiante pudo haber dormido bien antes de la fecha de una prueba pero en otro momento estaba demasiado ansioso para poder hacerlo. O en una ocasión hacía demasiado calor en la sala del examen pero no la siguiente vez. Otra fuente de error de medición son las anomalías en la calificación de las respuestas de los estudiantes.

A eso nos referimos por *confiabilidad*. Los resultados confiables muestran poca inconsistencia de una medición a la siguiente, es decir, contienen relativamente poco error de medición. Es común que se haga un uso incorrecto del término *confiabilidad* como sinónimo de "exacto" o "válido", cuando en realidad sólo se refiere a la *consistencia* de la medición. Una medida puede ser confiable pero inexacta, como una escala que consistentemente da lecturas demasiado altas. Estamos acostumbrados a mediciones de gran confiabilidad en muchos aspectos de nuestra vida, como cuando medimos la temperatura corporal o la longitud de una mesa que queremos comprar. Por desgracia, los resultados obtenidos en las pruebas educativas son mucho menos confiables que esas mediciones.

De modo que cuando todo está dicho y hecho, ¿qué tan justificable sería sacar conclusiones sobre el vocabulario a partir de la pequeña muestra de palabras evaluadas? Esta es la cuestión de la *validez*, que es el criterio individual más importante para evaluar las pruebas de logro. En el debate público, y a veces también en los estatutos y regulaciones, suele hacerse referencia a "pruebas válidas",

pero las pruebas en sí no son válidas o no válidas. Más bien, lo válido o no válido es la inferencia basada en las calificaciones. Una determinada prueba podría ofrecer buen apoyo para una inferencia pero no para otra. Por ejemplo, un examen final de estadística que esté bien diseñado podría sustentar las inferencias sobre el manejo que tienen los alumnos de la estadística básica, pero daría un apoyo muy débil a las conclusiones sobre la maestría más general de las matemáticas. La validez también es un continuo: es raro que la validez de las inferencias sea perfecta. La pregunta es *qué tanto apoyo* tiene la conclusión. Es difícil imaginar un ejemplo de una conclusión importante sobre el desempeño de un estudiante que fuese completamente apoyada por el rendimiento en una prueba, aunque no es difícil inventar uno que no tenga ningún apoyo en absoluto. Muchos de los problemas más específicos que se abordan en los capítulos posteriores –como la confiabilidad y el sesgo de la prueba– son piezas del rompecabezas de la validez.*

Nada de lo anterior es particularmente controvertido. Sin embargo el último paso en el ejemplo, es en verdad polémico. Suponga que es lo bastante amable para compartir conmigo su lista final de 40 palabras. O quizá, para ser más realista, no las comparte pero yo las veo o de alguna manera averiguo muchas de ellas. Suponga además que intercepto a cada aspirante que está en camino para presentar su evaluación y que le doy una breve lección sobre el significado de cada palabra de su examen. ¿Qué

* Los especialistas en medición suelen usar el término *validez* para referirse a los efectos de un programa de evaluación así como a la calidad de la inferencia basada en los resultados, y a menudo se refieren a esos resultados como "validez consecuente". Aunque sólo tengo elogios para la atención dirigida al impacto de la evaluación (he dedicado más tiempo de mi carrera a ese tema que la mayoría de mis pares), he descubierto que esa etiqueta por lo general confunde a la gente que no conoce a fondo la jerga. Por ende, como explico con mayor detalle en el capítulo 10, en este libro nunca empleo el término validez para referirme al impacto de la evaluación.

sucedería con la validez de las inferencias que usted desea fundamentar en los resultados obtenidos en su prueba?

Es claro que sus conclusiones acerca de qué aspirantes tienen vocabularios más ricos ahora serían erróneas. La mayoría de los estudiantes recibiría calificaciones perfectas, o casi perfectas, independientemente de sus verdaderos vocabularios. Los que prestaron atención durante mi mini-clase superarían a quienes no lo hicieron, aun si sus vocabularios fueran en realidad más pobres. El manejo de la pequeña muestra de 40 palabras ya no representa variaciones en los vocabularios reales de los alumnos.

Pero suponga que a usted no le interesa clasificar a sus candidatos en términos de la riqueza relativa de sus vocabularios: sólo quiere saber si sus vocabularios alcanzan un nivel que usted considera adecuado. No contratará a nadie cuyo vocabulario no alcance ese nivel "adecuado" y no le interesan las diferencias en el vocabulario de aquellos que alcancen o excedan ese nivel. En el oficio se dice que este último tipo de inferencias son *absolutas* porque no se compara el desempeño de un estudiante con el desempeño de otros sino con un estándar total. Muchos de los resultados más importantes de los actuales programas de evaluación en la educación elemental y media –incluyendo los que se utilizan para la rendición de cuentas de la ley «Que ningún niño se quede atrás»– son inferencias absolutas más que relativas. Por ejemplo, un artículo reciente en el *Washington Post* informaba que "en Virginia, los alumnos de cuarto grado mostraron pequeñas ganancias en matemáticas y lectura... 39 por ciento de los niños se consideran competentes en matemáticas, en comparación con 35 por ciento en 2003".[5] Además, la ley «Que ningún niño se quede atrás» exige que las escuelas sean recompensadas y castigadas con base en los cambios en los porcentajes de alumnos que alcancen ese nivel de desempeño. Para los propósitos de la rendición de cuentas no importa, y en ocasiones ni siquiera se reporta, qué tan

peor o mejor que "competente" es la calificación de un individuo o quiénes entre los estudiantes competentes tienen los niveles más altos de maestría.

Estas inferencias absolutas también se verían debilitadas si enseño a los aspirantes sus 40 palabras. Como resultado de mi pequeña lección, las 40 palabras ya no representan el dominio del vocabulario. En teoría, un estudiante podría no conocer otras palabras a excepción de las 40 y aún así obtendría una calificación perfecta. En principio, uno podría enseñar las 40 palabras, y nada más, a Koko la gorila (aunque en el lenguaje de señas), y también ella podría demostrar un vocabulario rico en la prueba. De modo que, una vez más, en esas condiciones (con mi lección a los aspirantes), el manejo de esa pequeña muestra no le diría nada útil acerca de la riqueza de los vocabularios de los aspirantes, es decir, si alcanzaron el umbral "adecuado".

Probemos esta vez con otra inferencia, también muy importante en los programas actuales de evaluación: que los vocabularios de los estudiantes mejoraron como resultado de mi clase. En la práctica real, este tipo de inferencias suele ser una defensa común para dedicar el tiempo instruccional a la preparación para la prueba: la gente argumenta que aunque es posible que los estudiantes no estén aprendiendo todo lo que usted quiere que aprendan, por lo menos están aprendiendo algo útil. En el caso de nuestro ejemplo, también esto sería aferrarse a una esperanza. Usted empezó por elegir palabras de dificultad moderada. Supongamos entonces que el aspirante promedio conocía más o menos la mitad de las palabras de su prueba. Cuando yo terminé con ellos, conocían las 40 palabras, al menos durante unos días, hasta que olvidaron algunas de ellas. Así que según un cálculo aproximado, sus vocabularios se habrían incrementado en 20 palabras de, digamos, 11,000 a 11,020 o de 17,000 a 17,020 palabras. Tal vez eso sea una mejora, pero difícilmente es suficiente para ameritar un comentario.

Puede haber casos en que el aprendizaje del contenido específico de la prueba constituya una mejora considerable, pero se mantiene la conclusión general de que incluso las inferencias sobre la mejora son debilitadas por ciertos tipos de preparación que se enfocan en la muestra concreta que se incluye en la prueba.

Este último paso en el ejemplo (enseñar a los estudiantes el contenido específico de la prueba o un material lo bastante cercano para socavar su representatividad) ilustra el polémico tema de la *inflación de las calificaciones o los resultados*, que se refiere a los aumentos en las notas que no señalan un incremento acorde en la competencia en el dominio de interés. Mi preparación para la prueba habría debilitado la validez de las distintas inferencias sobre el vocabulario que pudieran basarse en su examen hipotético, salvo la inferencia en esencia inútil sobre el manejo de las 40 palabras incluidas. En este caso, para que los resultados fueran inflados no fue necesario que la prueba tuviese algún defecto ni que se enfocara en un material poco importante. Las 40 palabras estaban bien, no así mi respuesta a esas 40 palabras (mi forma de preparación para la prueba). Lo que importa es la inferencia de la muestra examinada respecto a la competencia en el dominio, y cualquier forma de preparación que debilite esa conexión mina la validez de las conclusiones basadas en los resultados. En los programas reales de evaluación, los problemas de la inflación de las calificaciones y la preparación para la prueba son mucho más complejos de lo que sugiere este ejemplo, y en un capítulo posterior los tocaré de nuevo para mostrar la gravedad que pueden alcanzar y para explicar algunos de los mecanismos que les subyacen. ■

El ABC
de la
evaluación educativa

Capítulo 3
Lo que medimos: ¿qué tan buena es la muestra?

En el capítulo anterior me referí de pasada a una importante limitación de las pruebas de logro: "De la misma forma en que un encuestador no puede medir directamente la conducta futura de votar, hay algunos aspectos de las metas educativas que las pruebas de logro no pueden medir". Esta afirmación solivianta a muchos críticos de la educación y a menudo consigue que se etiquete a quienes la pronuncian como personas que están en contra de la evaluación o de la rendición de cuentas. Su argumento dice que las pruebas miden lo que es importante y que quienes se enfocan en otras "metas" son blandos.

Esas críticas no son del todo equivocadas. Algunas de las personas que hacen este reclamo sobre las limitaciones de la evaluación en efecto son contrarias a las pruebas estandarizadas y muchas se oponen a que se imponga desde fuera a las escuelas la obligación de rendir cuentas. Pero esto equivale a desviar la atención del verdadero problema. No es necesario oponerse a las pruebas o la rendición de cuentas en la educación (yo no lo hago) para reconocer sus limitaciones, e ignorarlas es una invitación para los problemas.

Cuando Richard Nixon inició sus acercamientos hacia China, los expertos se basaron sólo en sus antecedentes conservadores en la política exterior para afirmar de manera casi unánime que tenía la capacidad política de hacerlo. Según una lógica similar, una postura ventajosa para examinar esta limitación de la evaluación es un trabajo poco conocido en la actualidad que fue publicado hace

más de medio siglo por E. F. Lindquist de la Universidad de Iowa con el poco atractivo título de "Consideraciones preliminares en la elaboración de pruebas objetivas".[1] Lo que sea que uno pudiera decir de Lindquist, nadie podría acusarlo de estar en contra de la evaluación. De hecho, sería difícil pensar en alguien que hubiera hecho más que él para fomentar el desarrollo y uso de las pruebas estandarizadas de logro. Lindquist desarrolló en el campo la totalidad de su prolífica carrera y se le considera uno de los progenitores de la evaluación del logro en Estados Unidos. Aunque muy pocos fuera de la profesión reconocen su nombre, algunos de los productos de su trabajo son términos muy conocidos. Participó en el desarrollo de las Pruebas de Habilidades Básicas de Iowa (*Iowa Tests of Basic Skills*, ITBS), uno de los más antiguos instrumentos estandarizados de logro para la educación primaria y secundaria; las Pruebas de Desarrollo Educativo de Iowa (*Iowa Tests of Educational Development*), una batería de pruebas de logro para bachillerato de uso menos común; la prueba ACT de admisión a la universidad; la prueba GED de equivalencia para la graduación de bachillerato, y la Prueba Nacional de Cualificación del Mérito Académico (*National Merit Scholarship Qualifying Test*). Además, junto con sus colegas inventó el primer escáner óptico para la calificación de pruebas, una innovación que aceleró enormemente esa tarea y de esa forma contribuyó a la expansión de la evaluación estandarizada en las décadas de los cincuenta y los sesenta. Todo esto para decir que al considerar los argumentos de Lindquist podemos dejar a un lado la etiqueta de que estaba "en contra de la evaluación".

El trabajo de Lindquist es un buen punto de partida también por otras razones. Ahí ofreció una de las mejores explicaciones existentes de la evaluación estandarizada y mostró una notable clarividencia al anticipar las controversias que envolvieron al mundo de la política educativa décadas después de su trabajo. Eso puede hacer que el trabajo de Lindquist resulte desconcertante para

quienes se han enfrascado recientemente en la evaluación. Uno de mis alumnos, un antiguo maestro que había pensado mucho en la evaluación durante varios años antes de iniciar su posgrado, leyó el trabajo de Lindquist al final de mi clase y luego me envió un correo electrónico donde escribió: "1) ¿Por qué no había yo leído esto antes? 2) No sé si esto debe alegrarme o desalentarme. 3) Esto fue escrito 20 años antes de que yo naciera. ¿Queda algo que yo pueda hacer? 4) ¿Nadie lo ha leído? Me parece difícil creer que hayamos hecho algún progreso real en las áreas de las que él escribe". Su esposa agregó que se necesitaba un blog para Lindquist.

Lindquist planteó precisamente el argumento con el que inicié el capítulo: que las metas de la educación son diversas y que sólo algunas de ellas pueden someterse a la evaluación estandarizada.

Primero, dijo que si bien es fácil que hagamos evaluaciones para averiguar si un estudiante adquirió ciertos tipos de conocimiento y algunas habilidades particulares, es mucho más difícil evaluar otros tipos de destrezas. Un ejemplo actual e importante puede ser la habilidad para diseñar y llevar a cabo un experimento científico. Más difícil aún es medir algunas de las disposiciones y capacidades que muchos de nosotros deseamos que las escuelas fomenten, como el interés por el aprendizaje (los estudiantes necesitarán seguir aprendiendo durante toda su vida), la capacidad de aplicar de manera productiva el conocimiento adquirido en la escuela al trabajo posterior, etcétera.

Esto no implica, como le harían creer algunos críticos de la evaluación, que las pruebas estandarizadas sólo pueden medir cosas de poca importancia relativa; la evidencia muestra con claridad que pueden medir muchas cosas valiosas. Es evidente que Lindquist lo creía, pero al mismo tiempo advirtió que, sin importar lo valiosa que sea la información de una prueba de logro, sigue siendo necesariamente incompleta y que una parte de lo que omite es muy importante. Durante el medio siglo transcurrido desde entonces,

otros especialistas han repetido muchas veces esta advertencia. Por ejemplo, un manual reciente de la ITBS recomienda a los administradores escolares que traten explícitamente los resultados de las pruebas como información especializada que complementa –no sustituye– otras informaciones sobre el desempeño del estudiante. Por la misma razón, indica que es inapropiado usar los resultados de una sola prueba, sin información adicional, para asignar a estudiantes a educación especial, retenerlos, examinarlos para su primera inscripción, evaluar la eficacia de todo un sistema educativo o identificar a los "mejores" maestros o escuelas.[2] Una vez más, esta no es la postura de quienes se oponen a la evaluación, es el consejo de uno de los autores de las pruebas de logro más conocidas en Estados Unidos. Por desgracia, se ha ignorado en buena medida la advertencia de que las calificaciones o resultados de las pruebas, por muy útiles que sean, proporcionan información limitada; como veremos, esto se ha vuelto más pronunciado en los años recientes, a medida que los resultados de las pruebas se han convertido por sí mismos en un resumen de la medición de logro de los estudiantes y desempeño de las escuelas.

Segundo, Lindquist sostenía que incluso muchas de las metas de la educación que pueden someterse a las pruebas estandarizadas sólo pueden evaluarse de una forma menos directa de lo que nos gustaría. Muchos de los objetivos en que maestros y alumnos concentran cada día su atención son sólo sustitutos de las metas últimas de la enseñanza, demasiado generales y muy alejadas de las decisiones que deben tomarse continuamente en el aula. Por ejemplo, ¿por qué debería enseñarse álgebra a los estudiantes? Una razón, al menos desde mi punto de vista, es enseñarles a razonar de manera algebraica de modo que puedan aplicar este razonamiento a circunstancias relevantes fuera de la escuela. Sin embargo, este tipo de meta es demasiado general y se aleja mucho de las decisiones sobre el contenido de álgebra que debe enseñarse el

jueves por la mañana en una determinada escuela secundaria. Una vez que se ha decidido que los estudiantes deben estudiar álgebra, los maestros y los encargados de diseñar el currículo deben tomar muchas decisiones específicas acerca de qué álgebra enseñar. Por ejemplo ¿deben los estudiantes aprender a factorizar ecuaciones cuadráticas? Muchas consideraciones dan forma a esas decisiones, no sólo la posible utilidad que tenga el tema muchos años más tarde en una amplia variedad de contextos además de los relacionados con el trabajo.

Una anécdota puede aclarar la diferencia entre el aprendizaje del contenido especificado en un currículo y la aplicación posterior de dicho conocimiento. Hace muchos años, tuve un desayuno dominical en Manhattan con tres neoyorquinos. Los tres habían cursado la educación superior y habían llevado al menos dos semestres de matemáticas después de la preparatoria. Según mi experiencia, para orientarse en la ciudad los neoyorquinos recurren a un enorme conocimiento de la ubicación de edificios, como el de la librería original de Barnes y Noble en la Quinta Avenida. Ese domingo por la mañana, descubrí, para mi sorpresa, que ninguno de ellos podía discernir la ubicación del restaurante donde íbamos a desayunar. Se encontraba en una de las avenidas principales y ellos conocían la dirección, pero ninguno sabía qué calle la cruzaba. Sugerí que el problema podría ser muy sencillo. Pregunté si sabían dónde alcanzaban el cero las numeraciones de las avenidas en esa parte de Manhattan y, de saberlo, si lo hacían en la misma calle. Se mostraron de acuerdo en que así era y me dieron el nombre de la calle de cruce. Luego pregunté si la numeración aumentaba a la misma tasa en esas avenidas y, de ser así, a qué tasa. Es decir, ¿en cuántos números se incrementaba la numeración en cada cruce de calles? Se mostraron muy seguros de que lo hacían a la misma tasa, pero les llevó un poco más de trabajo averiguarlo. Utilizaron como referencia algunos edificios famosos (como el de

la librería original de Barnes y Noble) para calcular la tasa de un par de avenidas. Las tasas eran las mismas. En ese punto ya tenían la respuesta, aunque todavía no se habían dado cuenta. El problema era una simple ecuación lineal de una variable, como las que habían estudiado en secundaria, es decir, $y = a + bx$, y ya habían calculado *a*, el intercepto (la calle de cruce donde las direcciones alcanzaban el cero), y *b* (la pendiente, la tasa a la que se incrementaban las numeraciones). Como puede imaginar, se sorprendieron un poco cuando se los expliqué y les di la solución. Los tres eran competentes en el manejo de temas de álgebra mucho más complejos, pero no habían desarrollado el hábito de pensar en los problemas reales en términos de las matemáticas que aprendieron en el aula.

Ahora bien, es difícil argumentar que la capacidad de usar el álgebra para localizar un restaurante para el desayuno del domingo es el tipo de resultado educativo por el cual deberíamos perder el sueño, pero sí es seguro que tanto la habilidad como la inclinación a aplicar el conocimiento y las habilidades aprendidas en la escuela a empeños posteriores constituyen una meta importante. No hacemos que los estudiantes asistan a la escuela sólo para que les vaya bien mientras están ahí. Lo hacemos porque pensamos que los beneficia a ellos y a la sociedad como un todo, que los ayudará a tener más éxito en sus estudios posteriores, a ser más exitosos en el mundo del trabajo y mejores ciudadanos, y que les permitirá manifestar su potencial y llevar una vida más plena.

Por lo tanto –argumentaba Lindquist–, en un mundo ideal mediríamos las metas últimas de la educación para evaluar el logro. Escribió: "La única medida perfectamente válida del logro de un objetivo educativo sería una que se basara en la observación directa de la conducta natural de los... individuos... La medición directa es la que se basa en una muestra de la conducta natural, o criterio... para cada individuo".[3] Así, por ejemplo, si queremos saber si las escuelas tuvieron éxito al enseñar las habilidades y disposiciones

necesarias para usar el álgebra satisfactoriamente en el trabajo posterior, observaríamos a los estudiantes más tarde en la vida para ver si usan el álgebra cuando es pertinente y si lo hacen de manera adecuada.

Pero Lindquist sostenía que, por muchas razones, era claro que este tipo de medición no es práctico. En primer lugar, el criterio es demorado. No podemos permitirnos esperar una o dos décadas para saber si los alumnos de octavo grado de este año hacen uso del álgebra en su trabajo adulto. Incluso si pudiésemos hacerlo, las conductas criterio (en este caso, la aplicación adecuada del álgebra cuando es apropiado) suelen ser infrecuentes. Yo uso el álgebra muy a menudo, pero es probable que la mayoría de ustedes no lo haga, por lo que un observador tendría que vigilarlos por un periodo muy largo para saber si adquirieron esas habilidades y disposiciones. E incluso si estuviésemos dispuestos a esperar que se presentaran esos resultados, esta forma de medir el logro resultaría demasiado costosa en tiempo y esfuerzo. Más aún, algunos de los criterios (por ejemplo, algunas disposiciones) no pueden observarse directamente.

Esas razones parecen obvias, pero Lindquist añadió otras dos que si bien son menos evidentes, tienen profundas implicaciones para la evaluación y dieron lugar a la forma particular que adoptaron la ITBS y la mayor parte de las pruebas estandarizadas de logro de la época. Primero, señaló que las muestras de conducta que ocurren de manera natural no son comparables. Por ejemplo, suponga que yo utilicé un poco de álgebra esta mañana mientras que un amigo mío, director de una escuela de leyes, no lo hizo. ¿Significa eso que yo logré adquirir más de este conjunto de disposiciones y habilidades que el director de la escuela de leyes? La conjetura no es irracional; después de todo, la gente a la que no le gustan o que no es muy buena con las matemáticas es mucho menos proclive a elegir la psicometría que, digamos, el derecho como profesión.

Pero aunque esto parezca una conjetura sensata, no es una suposición segura, y en este caso es del todo errónea: el director de la escuela de leyes se especializó en matemáticas durante su licenciatura en una escuela particularmente exigente. ¿Cómo explicamos entonces el hecho hipotético de que yo utilicé el álgebra y él no lo hizo? Cada uno de nosotros enfrenta demandas diferentes en el trabajo: yo necesito usar el álgebra a menudo mientras que él rara vez lo requiere. A partir de la simple observación de cada uno de nosotros en el trabajo, no se sabría con certeza cuál de varias explicaciones es correcta; es decir, si el director de la escuela de leyes utiliza menos el álgebra porque no siente la inclinación a hacerlo, porque no lo hace bien o simplemente porque su trabajo no lo requiere. En términos de Lindquist, nuestros ambientes laborales no son comparables y, por ende, nuestra conducta (si usamos o no el álgebra) no necesariamente significa lo mismo sobre el logro de cada uno de nosotros.

Segundo, Lindquist advirtió que algunas conductas criterio son complejas y requieren diversas habilidades y conocimientos. En tales casos, si una persona tiene un mal desempeño en una de esas conductas criterio, no se sabría la razón. De las distintas cosas que es necesario saber para realizar bien la tarea, ¿cuál de ellas no entiende una persona sin éxito? De las diversas habilidades necesarias, ¿de cuál carece la persona? Y, lo que resulta de primordial importancia para la idea de Lindquist sobre el papel de la evaluación del logro: si no es posible identificar la razón del mal desempeño ¿cómo puede mejorarse la instrucción en consecuencia?

Lindquist explicó luego en detalle algunas implicaciones de este razonamiento para la evaluación del logro. En primer lugar, el autor de una prueba por lo general debe enfocarse en las metas próximas de los educadores, aunque sólo sean sustitutas de las metas sociales últimas de la educación. En opinión de Lindquist, esto significa concentrarse sobre todo en el currículo. En la jerga actual, esto se

traduce como "alineación de la prueba con los estándares", pero la idea básica es la misma. En segundo lugar, dado que no podemos esperar por años para ver si ocurre una conducta, tenemos que hacer algo para lograr que se presente ahora. Ese algo es la prueba. En tercero, para evitar el problema de confundir las diferencias de conocimiento y habilidades con las diferencias en los ambientes de las personas (el mío y el de mi amigo, el director de la escuela de leyes), tenemos que poner a todos los examinados en el mismo ambiente cuando provoquemos las conductas que vamos a medir. Ello significa que tenemos que *estandarizar* la prueba y hacer que sea la misma para todos los estudiantes.

En gran medida esto es relativamente poco polémico. Es verdad que el término *prueba estandarizada* suele usarse con un tono desdeñoso en el debate actual, pero como expliqué en el capítulo anterior, es común que eso refleje una mala interpretación del término más que una objeción a la estandarización. Casi todas las pruebas de logro de gran escala son estandarizadas, sin importar lo que incluyan ni la forma que adopten, y la mayoría de los maestros intentan estandarizar sus evaluaciones en el aula para sus alumnos.

No obstante, otras dos implicaciones del razonamiento de Lindquist son controvertidas y, de hecho, se han vuelto más polémicas en las décadas recientes. Primero, para ayudar a orientar la instrucción, Lindquist pretendía nada menos que aislar el conocimiento y las habilidades específicas. En su opinión, esto requería que las pruebas se diseñaran para incluir tareas que se enfocaran de cerca en esos detalles. Por ejemplo, Lindquist habría argumentado que si usted desea determinar si los estudiantes de tercer grado pueden manejar las restas con acarreo, debe aplicarles problemas que requieran ese tipo de restas pero que conlleven tan pocas habilidades secundarias como sea posible. No debería insertar esa habilidad en un texto complejo porque entonces un estudiante podría fracasar en la solución del problema por carecer de

esas habilidades aritméticas o porque no sabe leer, y sería difícil identificar la causa. Este principio se refleja todavía en el diseño de algunas pruebas, pero en otros casos los reformadores y los desarrolladores de las pruebas han pasado deliberadamente a la dirección opuesta en un intento por crear reactivos de prueba que presenten tareas complejas, "auténticas", que sean más parecidas a las que los estudiantes pueden encontrar fuera de la escuela. Ambos lados de este argumento tienen aciertos y errores: hay ventajas y desventajas en ambas maneras de diseñar las pruebas. Este es uno de los muchos casos en que el diseño de las pruebas implica compromisos e intercambios de metas que compiten entre sí.

Por último, Lindquist sostuvo que la interpretación del desempeño en las pruebas refleja el hecho necesario y sistemático de que son incompletas. Es decir, la calificación de una prueba debería considerarse como una medida de lo que pueden hacer los estudiantes en una parte particular e importante, pero limitada, de los resultados que se espera que las escuelas produzcan. Por lo tanto, lo ideal sería utilizar los resultados de las pruebas como complemento de otras informaciones sobre el desempeño de los alumnos. Esas otras informaciones tendrán fortalezas y debilidades distintas a las de los resultados de la prueba. Una de esas fuentes de información es lo que los maestros recogen en el curso de su enseñanza y en las evaluaciones en el aula. Un maestro sagaz puede observar muchas cosas que es difícil evaluar, pero sus juicios carecen de la estandarización que ofrecen las calificaciones de examen y, por lo tanto, son mucho menos comparables de un entorno a otro. Por ejemplo, sabemos que la calificación de los maestros es, en promedio, mucho más indulgente en las escuelas con mucha pobreza que en las que la pobreza es poca. Al integrar la información de varias fuentes con diferentes fortalezas y debilidades, podemos hacernos una idea más completa de lo que los estudiantes saben y pueden hacer.

Las opiniones de Lindquist sobre este último tema se reflejan todavía en las advertencias de algunos expertos en medición. Una vez más, el manual de la ITSB que mencionamos antes sirve como ejemplo: advierte a los administradores de no usar los resultados de las pruebas por sí solos como una evaluación definitiva de una escuela o programa.[4]

¿Realmente se sigue en la práctica el consejo de Lindquist? En algunas partes sí. Eso es justamente lo que hacen los encargados de los departamentos de admisión de las universidades cuando realizan una revisión "holística" de los candidatos, en que no sólo consideran los resultados de exámenes como el SAT o el ACT, sino también las notas escolares, las exposiciones personales, la persistencia en actividades extracurriculares, etcétera. Las universidades tienden a ser muy reservadas acerca de sus procesos de admisión; en alguna ocasión, una persona que durante mucho tiempo dirigió el departamento de admisión de una universidad muy selectiva (no, no era Harvard) me dijo que su política era revelar a cada solicitante potencial, tan poco como necesitaran saber acerca de lo que hacía su oficina. Sin embargo, puedo decirles que nuestra selección de candidatos en la Escuela de Posgrado en Educación de Harvard es congruente con el consejo de Lindquist. Los miembros de los comités de admisión en los que yo he participado consideran que cada medida individual, incluyendo los resultados del Examen de Registro de Posgrado (*Graduate Record Examination*, GRE), está muy lejos de ser completa o suficiente. Los comités han intentado integrar una visión general a partir de todos los datos de que disponen y a menudo debaten cómo deben interpretar las contradicciones entre ellos (como el hecho de que las notas escolares sean sensiblemente superiores a las calificaciones del examen o viceversa).

Por desgracia, esto suele ser más la excepción que la regla y, en gran parte de la evaluación que domina ahora la educación básica

y media, es común que se ignore el consejo de Lindquist de considerar que los resultados obtenidos en la prueba son medidas incompletas. Las pruebas más importantes se han convertido en un medio para supervisar a las escuelas y hacerlas responsables, lo que implica una desviación considerable del consejo de Lindquist. En la actualidad se usan los resultados de una sola prueba como si fueran un resumen exhaustivo de lo que los estudiantes saben o de lo que las escuelas producen. Es irónico y desafortunado que a medida que la evaluación ha aumentado su importancia para la educación estadounidense, nos hemos alejado más del inteligente consejo que nos ofreció hace mucho uno de los más importantes y eficaces defensores de la evaluación estandarizada en el país. ■

Capítulo 4
La evolución de la evaluación en Estados Unidos

A finales de la década de los noventa, los hijos de las familias que se mudaban a unos cuantos kilómetros del río Potomac, en cualquier dirección, enfrentaban en sus escuelas programas de evaluación totalmente distintos. El el Programa de Evaluación del Desempeño Escolar de Maryland (*Maryland School Performance Assessment Program*, MSPAP), en ese tiempo vigente en Maryland, sólo incluía grandes tareas de desempeño que requerían que los estudiantes escribieran sus respuestas, algunas preveían actividades de grupo, manipulación de aparatos y trabajo realizado durante varios días. Al otro lado del río, las escuelas de Virginia usaban pruebas de opción múltiple. Las explicaciones de lo que esas pruebas estaban diseñadas para medir (que en Virginia se denominaban estándares de aprendizaje y en Maryland resultados de aprendizaje) también eran sorprendentemente diferentes: los estándares de aprendizaje de Virginia tendían a ser mucho más detallados y específicos que los resultados de aprendizaje de Maryland. El contraste entre ambas ciudades era en verdad extremo, no así el patrón: existen variaciones evidentes en los detalles de los programas de evaluación de los estados e incluso, en menor grado, entre las localidades de algunos estados.

A pesar de esta diversidad, la mayor parte de los programas estadounidenses de evaluación a gran escala comparten varias características fundamentales. Casi todos son pruebas "externas", es decir, son exigidas por organismos externos a la escuela. Casi

todos los organismos estatales de educación imponen las pruebas a las escuelas públicas y muchos distritos locales añaden otras. Prácticamente todos los programas usan pruebas estandarizadas, lo que significa que su contenido, aplicación y calificación son, al menos desde el punto de vista ideal, uniformes de un niño a otro o de una escuela a otra. Muchas pruebas tienen el propósito de monitorear el desempeño de escuelas y estados completos y, en particular, de medir el cambio en su desempeño a lo largo del tiempo. Casi todas tienen un impacto considerable para los educadores, los estudiantes o ambos. Esas características de los programas de evaluación se han vuelto tan comunes que pocos le prestan mucha atención.

Pero la naturaleza de la evaluación educativa ha sufrido cambios considerables en las décadas recientes y muchos de los rasgos que ahora damos por sentado en realidad son bastante nuevos. Lo más importante, se ha dado un cambio fundamental en las funciones primordiales de la evaluación del logro de gran escala ya que la rendición de cuentas ha sustituido gradualmente al diagnóstico de las fortalezas y debilidades en el aprendizaje de estudiantes individuales. Este cambio en el uso de las pruebas ha sido acompañado por cambios en los tipos de conclusiones que se pretende apoyar con los resultados obtenidos en las pruebas. Las inferencias acerca de estudiantes individuales siguen siendo importantes –de hecho, en muchos estados y localidades esas conclusiones tienen consecuencias mucho más serias que hace tres o cuatro décadas– pero, en muchos casos, las conclusiones acerca del desempeño de grupos, en particular el desempeño de escuelas y distritos, son mucho más trascendentales. Las conclusiones sobre el logro en cualquier momento dado han cedido el paso a inferencias acerca de cambios en el desempeño a lo largo del tiempo; en particular, cambios en el desempeño conjunto de escuelas y distritos. En buena medida, los métodos tradicionales para reportar el desempeño en las pruebas

–muchos de los cuales comparaban el desempeño de un alumno con el de otros estudiantes– dieron paso a métodos que comparan los resultados de los educandos con las expectativas establecidas por los políticos u otros interesados. El formato de las pruebas a gran escala también se modificó y ahora depende menos de los tradicionales reactivos de opción múltiple.

El uso a gran escala de pruebas grupales de logro se remonta por lo menos a la década de los cuarenta del siglo XIX en Estados Unidos, y al menos un programa actual de evaluación, el programa *New York State Regents Examination*, data de la segunda mitad de ese siglo. Sin embargo, para nuestros propósitos es suficiente con retroceder apenas medio siglo, a la década de los cincuenta.

Los lectores de mi edad recordarán a los cincuenta como una era en que la evaluación tenía poco impacto, es decir, era raro que las pruebas tuviesen consecuencias graves. Pocos estados imponían los programas de evaluación, pero muchos distritos escolares compraban pruebas de logro de editores comerciales y las aplicaban anualmente. El distrito en que yo asistí a la escuela, Syracuse, Nueva York, aplicaba la Prueba de Habilidades Básicas de Iowa, que era una de las cinco pruebas comerciales que dominaban en esa época la evaluación del logro. Las cinco pruebas eran estandarizadas. En su mayor parte o en su totalidad eran de opción múltiple, lo que permitía evaluar una cantidad importante de contenido en un tiempo breve, mantener los costos bajos y una calificación automatizada perfectamente constante.

La Prueba de Habilidades Básicas de Iowa y otros instrumentos similares que en mi juventud dominaban la evaluación del logro (y que en la actualidad siguen siendo de gran uso, aunque con una competencia mucho mayor de pruebas más recientes que se han hecho populares) se diseñaron con un propósito primordial de diagnóstico: ayudar a los maestros y administradores a identificar las fortalezas y debilidades relativas en el logro de sus estudiantes.

También pretendían identificar las áreas de fortaleza y debilidad dentro de las escuelas y distritos escolares para facilitar la mejora de la educación. Sin embargo, no se pretendía que proporcionaran evaluaciones sucintas del desempeño de las escuelas, distritos, estados o países ni que hicieran responsables a los educadores.

Estas pruebas no se consideraban triviales, pero en la mayoría de los casos los estudiantes y los maestros no sufrían consecuencias graves ni cosechaban grandes recompensas como resultado de sus calificaciones. Había, sin embargo, algunas excepciones. En el caso de los estudiantes que competían por ser aceptados por universidades selectivas, las pruebas de admisión como el SAT (en ese entonces conocida como Prueba de Aptitud Escolar, por un tiempo Prueba de Evaluación Escolar y ahora sólo como SAT) tenían un impacto relativamente alto, aunque todavía no aparecía el frenesí generalizado de los cursos de preparación para esas pruebas que décadas más tarde envolvería a toda la cohorte de mis hijos. En algunas jurisdicciones, las pruebas de logro y otras, como las de CI, también eran propuestas de alto impacto para los estudiantes en la cúspide de la asignación a la educación especial u otras ubicaciones especiales. No obstante, para la mayoría de nosotros y de nuestros maestros, la evaluación del logro que encontrábamos antes de las pruebas de admisión a la universidad no generaba mucha ansiedad.

Por lo regular, esas pruebas descomponían el logro en muchos pedazos pequeños, cosa que todavía se hace en las ediciones actuales. Por ejemplo, una edición reciente de la Prueba de Habilidades Básicas de Iowa para estudiantes de secundaria descompuso las operaciones con fracciones, decimales y porcentajes en cuatro categorías diferentes de reactivos de examen que incluían "comparar y ordenar" y "aplicar razón y proporción a la solución de problemas".[1] Esta fragmentación del desempeño recientemente fue desdeñada por algunos críticos que argumentan a favor de incluir las

habilidades en tareas más grandes y más realistas; pero el propósito de dividir el desempeño en habilidades diferenciadas era sencillo. Era el argumento de Lindquist, explicado en el capítulo anterior: si se descomponen las habilidades y el conocimiento en pedazos pequeños, es más fácil discernir las habilidades específicas que contribuyen a los puntos flacos de los alumnos y ayudar de este modo a los educadores a mejorar su enseñanza. Por ejemplo, si podemos localizar con precisión las habilidades faltantes específicas que ocasionan que los estudiantes tengan un pobre desempeño en los problemas con fracciones, el maestro o el administrador escolar pueden mejorar la enseñanza de esas habilidades particulares.

Los instrumentos de mayor uso en la década de los cincuenta eran *pruebas referidas a normas* (PRN). Una prueba referida a normas es simplemente una en que el desempeño se reporta en comparación con la distribución de calificaciones o resultados de algún grupo de referencia. Los estándares para la comparación se denominan *normas* y el grupo del cual se obtuvieron las normas por lo general se conoce como "grupo normativo" o "muestra de estandarización". Por ejemplo, los padres de un estudiante al que le fue aplicada una prueba referida a normas podrían recibir un reporte como este: "El resultado de Mark en lectura lo coloca en el rango percentil 60 a nivel nacional", lo cual significa que 60 por ciento de los estudiantes de una muestra representativa del país obtuvo resultados menores a los de Mark.

El término *normas* suele causar confusión porque en economía su significado es justo el contrario. En ese campo la "economía normativa" se refiere a estudios que conllevan juicios evaluativos, mientras que a los estudios libres de valor se les asigna la etiqueta "economía positiva". En contraste, en medición las normas no tienen nada que ver con los valores; son sólo una distribución del desempeño utilizada como estándar de comparación para dar significado a los resultados. Fue justamente el hecho de que el reporte

referido a normas fuese libre de valor y puramente descriptivo lo que lo llevó a encontrar en años recientes la oposición de muchos educadores y políticos. Regresaré a este punto en el capítulo 8 cuando revisemos la alternativa, los estándares de desempeño.

Si bien el grupo normativo suele ser una muestra nacional de estudiantes, puede ser cualquier grupo útil de comparación. Por ejemplo, el promedio de una escuela puede compararse con una distribución nacional de promedios escolares, y el desempeño en un determinado tipo de escuela (digamos, escuelas católicas o escuelas que atienden a estudiantes de bajos ingresos) puede compararse con la distribución del desempeño en escuelas similares. Los grupos normativos pueden incluso constar de estados o países, y el reporte referido a normas puede utilizarse con cualesquier instrumento que muestre variaciones en el desempeño. Por ejemplo, las comparaciones internacionales periódicas del desempeño de los estudiantes que reciben gran difusión en la prensa, como el Estudio Internacional de las Tendencias en Matemáticas y Ciencia (*Trends in International Mathematics and Science Study*, TIMSS), emplean normas nacionales; el nivel de desempeño de cada país es aclarado comparándolo con la distribución del desempeño en otros países participantes. En ausencia de dichas normas es difícil saber qué hacer con el resultado promedio de Estados Unidos que es de 504 para matemáticas de octavo grado. Saber que los resultados promedio fueron similares en Austria, Nueva Zelanda, Escocia y Suecia, pero mucho más altos en Corea, Taiwán y Singapur, nos ayuda a entender los resultados de Estados Unidos.[2] (En el siguiente capítulo se describen algunos hallazgos importantes de los estudios internacionales.)

Hacemos un uso rutinario del reporte referido a normas, a menudo sin prestarle atención, para entender todo tipo de información en nuestras vidas. Para saber si mi carro (que recorre alrededor de 32 millas por galón en la carretera) es eficiente lo

comparo con la camioneta de mi vecino o con el Prius de un buen amigo, que recorren más o menos la mitad de lo que recorre mi auto, o bien lo comparo con las normas más amplias que se encuentran en los *Reportes del consumidor* o en los cálculos de EPA. ¿Cómo sabemos si un estudiante de preparatoria que corre una milla en poco más de cuatro minutos es una estrella? De nuevo recurrimos a las normas. ¿Cómo sabemos si debemos compadecer a una colega que nos dice que su viaje de Washington a Chicago le llevó ocho horas de puerta a puerta? Porque por muy malos que se hayan vuelto los viajes aéreos, sabemos que ocho horas es un tiempo inusualmente largo. ¿Por qué los periodistas siempre logran que sus editores les concedan espacio para historias acerca de evaluaciones internacionales como la prueba TIMSS? Porque las normas nacionales (en particular, nuestra posición en comparación con la de los países de mayor desempeño como Singapur y Japón) dan a los lectores una manera de juzgar la competitividad del desempeño estadounidense que les sería muy difícil evaluar de otra forma. (Resulta interesante preguntar por qué parece de interés periodístico el hecho de que califiquemos más bajo que Singapur pero no que igualemos a Australia, Suecia e Inglaterra y que superemos a Noruega –resultados obtenidos en este caso de la repetición en 2003 de la TIMSS–, pero es una pregunta para la cual no tengo respuesta.) En cada caso, utilizamos normas para entender la información cuantitativa que de otra manera resultaría difícil interpretar.

No obstante, en muchas partes el reporte referido a normas del desempeño en las pruebas tiene una mala fama inmerecida. Al parecer, existen tres razones para ello. La primera es el simple desconocimiento. Mucha gente cree equivocadamente que "referido a normas" alude a algo distinto a la manera en que se presentan los resultados; por ejemplo, a algo acerca del contenido o el formato de la prueba (opción múltiple). Si les disgusta algo de esas otras

cosas; por ejemplo, si les parecen objetables los exámenes de opción múltiple, en lugar de referirse a ellos critican erradamente a la evaluación referida a normas.

La segunda razón para criticar la evaluación referida a normas es la idea de que ese tipo de reporte, en lugar de presentar una exposición clara de si el desempeño está a la altura de las expectativas de alguien, predispone a la gente a aceptar el *statu quo*. "Por arriba del promedio" suena bien –prosigue el argumento–, pero si el desempeño promedio está muy por debajo de las expectativas, un desempeño que lo supere puede aún así ser inaceptable. Esto es una simplificación excesiva. Es verdad que el reporte referido a normas no necesariamente ofrece información acerca de si el desempeño alcanzó expectativas razonables, pero como muestran los ejemplos presentados antes, suele ser útil (de hecho algunas veces es esencial) precisamente para ese propósito. Más aún, el informe referido a normas puede asociarse con otras formas de reporte que comparen directamente el desempeño con las expectativas, como el reporte basado en estándares (que se revisa más adelante). En la actualidad, casi todos los estados reportan el desempeño de los alumnos en comparación con expectativas. Por ejemplo, los reportes públicos suelen mostrar el porcentaje de alumnos de una escuela que alcanzan un estándar de desempeño denominado "competente". Pero debido a su utilidad, es común que también se informe de los resultados referidos a normas: por ejemplo, se compara el porcentaje que alcanza el estándar "competente" en una determinada escuela con el porcentaje de todo el estado, de modo que los padres y la prensa puedan ver si el desempeño de la escuela es atípicamente alto o bajo.

El tercer argumento en contra de la evaluación referida a normas refleja un malentendido generalizado: que la evaluación referida a normas en realidad contribuye a las variaciones en el desempeño de los estudiantes. No siempre queda claro si los críticos pretenden

advertir que la evaluación referida a normas crea una ilusión de variación o que exagera la variación que en realidad existe. Durante un segmento del programa *Talk of the Nation* de la Radio Pública Nacional transmitido en marzo del 2002 desde la Escuela del Posgrado de Educación de Harvard, Jeff Howard, un psicólogo social que es un destacado defensor de los estándares de la educación superior, hizo la siguiente declaración: "La evaluación referida a normas, con la que crecimos muchos de nosotros, crea ganadores y perdedores. Tú estás en el decil más alto, tú en el decil inferior, tú estás en el promedio y tú estás por arriba del promedio. Están diseñadas para designar ganadores y perdedores".[3] Este es uno de los mayores errores en el debate actual sobre la evaluación. Las pruebas quizá "designen" a los ganadores y los perdedores, pero no los *crean*. Sencillamente *son* ganadores y perdedores. En cualquier lugar del mundo donde se mire, incluso en las sociedades más equitativas, es enorme la variación en el desempeño de los estudiantes. Por ejemplo, si revisa las pruebas de matemáticas de octavo grado, encontrará que la variación total del logro de los estudiantes estadounidenses es similar a la que existe en Japón, una sociedad socialmente más homogénea en que no hay formación de grupos según la capacidad hasta octavo grado.[4] Si uno decide no medir esa variación no la verá, pero aún así existe. Ya vimos esto en el capítulo 2: las pruebas de vocabulario no *causan* que algunos estudiantes conozcan menos palabras que otros. Sólo ayudan a establecer quién conoce más y quién menos. Para otros propósitos uno puede decidir que no se clasificará a los estudiantes y, por lo tanto, no se les aplicará una prueba diseñada para mostrar todas las variaciones en el desempeño, pero esas variaciones seguirán existiendo. El árbol hace ruido cuando cae, incluso si no hay nadie para escucharlo.

De igual modo, cuando los críticos usan "prueba estandarizada" como término de oprobio –cosa que suelen hacer– no

siempre queda claro de qué se quejan. El término se utiliza mal para denotar toda clase de cosas: pruebas referidas a normas; pruebas que incluyen tipos específicos de tareas (en particular, pruebas de opción múltiple); pruebas externas impuestas a las escuelas por organismos externos, y pruebas desarrolladas por editores comerciales u otros para una audiencia nacional, no para estados o distritos individuales. Esos usos del término sólo desvían la atención del verdadero problema. Es posible, en efecto, encontrar pruebas estandarizadas con esos atributos; por ejemplo, la Prueba de Habilidades Básicas de Iowa sigue siendo una prueba de opción múltiple, externa, referida a normas, desarrollada por un grupo de investigación universitario y comercializada nacionalmente por un editor importante. Sin embargo, una prueba puede ser estandarizada sin tener ninguno de esos atributos. Una prueba que requiere que los alumnos realicen experimentos científicos puede ser estandarizada, lo mismo que una evaluación que se reporte en términos de expectativas del desempeño en lugar de normas, siempre que las tareas, aplicación y calificación sean uniformes. De hecho, casi todos los programas a gran escala de evaluación del logro o admisión que se utilizaron en Estados Unidos durante el pasado medio siglo han sido pruebas estandarizadas. Una excepción en los programas de evaluación de gran escala son las llamadas evaluaciones de portafolios, en que los alumnos compilan para su evaluación posterior una colección de productos generados durante el curso del trabajo de un salón de clases regular. Unos cuantos estados, como Vermont y Kentucky, pusieron en práctica la evaluación por portafolios en las décadas recientes. En esos programas no se estandarizaron las tareas ni las condiciones administrativas en que se producía el trabajo, aunque sí la calificación.

El sistema común de la evaluación del logro de bajo impacto, diagnóstica y referida a normas empezó a cambiar durante la década de los sesenta, aunque al inicio muy lentamente. Hasta

entonces, el uso de los resultados de las pruebas para supervisar el desempeño de los sistemas educativos se limitaba, en su mayor parte, a los distritos locales que decidían vigilar la eficacia de sus propias escuelas. En los años sesenta, dos acciones del gobierno federal empezaron a cambiar esta situación. En 1965, el Congreso aprobó la Ley de Educación Primaria y Secundaria (*Elementary and Secondary Education Act*, ESEA), la cual establecía el Título I del programa de educación compensatoria –precursor del programa «Que Ningún Niño se Quede Atrás»– para mejorar el desempeño de los alumnos de escuelas de bajos ingresos. Esto marcó la primera participación importante del gobierno federal en el financiamiento y dirección de la educación general primaria y secundaria. La ley también exigía la evaluación del programa del Título I (la primera vez que la legislación federal estableció un programa social importante que requería un programa formal de evaluación). En 1974 el Congreso estableció el Sistema de Evaluación y Reporte del Título I (*Title I Evaluation and Reporting System, TIERS*), que basó las evaluaciones requeridas de los programas del Título I en los resultados obtenidos por los estudiantes en pruebas de logro estandarizadas y referidas a normas.

En una acción ocurrida a finales de la década de los sesenta, el gobierno federal estableció la Evaluación Nacional del Progreso Educativo (*National Assessment of Educational Progress*, NAEP), una evaluación periódica de muestras representativas de estudiantes de todo el país. Desde su inicio, este programa ha evolucionado en muchos sentidos; por ejemplo, en 1990 se modificó para permitir comparaciones frecuentes entre los estados (una forma de reporte referido a normas) que ahora reciben mucha atención de la prensa. Sin embargo, la Evaluación Nacional del Progreso Educativo fue desde el principio mucho más mesurada. Aunque brindaba información acerca del país como un todo, de regiones y de los principales subgrupos de la población joven (como los grupos

raciales), de manera deliberada se diseñó para que no pudiera proporcionar datos estatales o el distritales, donde se toma la mayor parte de las decisiones, por lo que no era útil para la rendición de cuentas. Su único propósito era ofrecer al público y a los tomadores de decisiones una descripción del logro de los estudiantes e información acerca de las tendencias del desempeño a lo largo del tiempo.

Ni NAEP ni TIERS imponían sus consecuencias en los estudiantes o los maestros a partir de los resultados de la prueba. No obstante, en retrospectiva parece que esos dos programas federales marcaron el inicio de un cambio radical en la evaluación educativa estadounidense. Significaron el comienzo de un cambio fundamental en las metas de la evaluación, del diagnóstico y la evaluación local al monitoreo en gran escala del desempeño y, por último, a la rendición de cuentas basada en las pruebas.[5]

No tardó mucho en darse el siguiente paso en esta evolución: el movimiento de evaluación de competencias mínimas. El primer programa estatal de evaluación de competencias mínimas se estableció en 1971; para finales de la década existían en vigor programas similares en 35 estados. El diseño de esta forma de evaluación pretendía asegurar que todos los estudiantes alcanzaran un nivel mínimo aceptable de dominio de habilidades básicas. Casi todas las pruebas de competencia mínima eran exámenes de salida en que los estudiantes tenían que superar una calificación especificada, llamada *punto de corte*, para obtener el diploma de preparatoria. Un número menor de programas utilizó los puntos de corte de esas pruebas como "puertas para la promoción", es decir, como requisitos que los alumnos tenían que cumplir para la promoción entre grados.[6] Como su nombre lo implica, esas pruebas por lo regular eran sencillas y sus calificaciones, o puntos de corte, bajas. Aunque todavía existen algunos de esos programas (por ejemplo, mientras escribía este libro, la ciudad de Nueva York y la

de Chicago usaban pruebas como "puertas para la promoción") durante la década de los ochenta declinó el uso de la evaluación estatal de competencia mínima.

A pesar de su corta duración, el movimiento de evaluación de competencias mínimas tuvo al menos cuatro efectos importantes y duraderos en la evaluación del logro a gran escala en Estados Unidos. Primero, y quizá lo más evidente, fue otro paso importante en dirección al uso de las pruebas para la rendición de cuentas. Por primera vez desde la Segunda Guerra Mundial, en la mayoría de los estados, se hizo a los estudiantes directamente responsables de su desempeño en una prueba, y si bien los estándares eran bastante bajos para que sólo un porcentaje modesto de ellos reprobara, las consecuencias para quienes lo hacían eran graves. Segundo, el movimiento dio inicio a un incremento notable en el número de estados con programas de evaluación obligatorios en toda la entidad, un crecimiento que continuó después de que las pruebas de competencia mínima perdieron popularidad. Aunque algunos estados, como Nueva York, tenían desde hacía mucho tiempo programas de evaluación, antes de 1970 no era así en la mayoría de los estados. Para finales de la década de los setenta, 60 por ciento de las entidades tenían programas de evaluación obligatorios y 20 años más tarde casi todos los tenían.

Tercero, el movimiento de evaluación de competencia mínima señaló el inicio de un cambio hacia el reporte del desempeño del estudiante en comparación con las expectativas más que con las normas. Esos programas de evaluación por lo general empleaban pruebas referidas a criterio (PRC) y reportaban el desempeño en términos de qué tan bien había dominado cada estudiante algún conjunto establecido de conocimientos y habilidades (el criterio), sin compararlo con otros estudiantes. En la práctica, los estados y localidades que emplean las pruebas referidas a criterio por lo general establecen una calificación o punto de corte para el desempeño,

una sola calificación que constituye el nivel aprobatorio en la prueba, como las calificaciones o resultados en las pruebas de competencia mínima requeridas para la graduación. Desde entonces el reporte referido a normas ha ido y venido, pero la evaluación referida a criterio con puntos de corte se ha mantenido y, como señalo más adelante, ahora es exigida por la ley federal, aunque con un nombre diferente.

Menos obvio pero de mayor importancia, el cuarto efecto duradero de la evaluación de competencia mínima fue un cambio fundamental en la forma en que se usan las pruebas para mejorar la instrucción. La idea que está detrás de las pruebas tradicionales de logro, como la Prueba de Habilidades Básicas de Iowa, es que las pruebas estandarizadas deben mejorar la instrucción al proporcionar a los educadores y a los padres información útil de la que de otro modo carecerían. Pero la expectativa que motivó el movimiento de evaluación de competencia mínima era que la instrucción mejoraría si se hacía a alguien (en este caso los estudiantes, pero también podían ser los educadores) directamente responsable del desempeño en las pruebas. Esta idea aparentemente llena de sentido común recibió el nombre de "instrucción dirigida por la medición".

No hay duda de que la modificación en el uso de las pruebas –de usarse para obtener información a emplearse para hacer a los alumnos o a los maestros directamente responsables de las calificaciones– fue el cambio individual más importante en la evaluación en el pasado medio siglo. La rendición de cuentas basada en la evaluación ha tomado diversas formas de un lugar a otro y de un momento a otro en los pasados 30 años, pero el principio básico de moldear la práctica educativa por medio de la responsabilidad por los resultados obtenidos en las pruebas se ha vuelto más importante para la política educativa en Estados Unidos (y en muchos otros países). No es una exageración decir que ahora es la

piedra angular de la política educativa estadounidense. Esta tendencia culminó en la promulgación en el 2001 de la ley «Que ningún niño se quede atrás», pero faltan por documentar algunos pasos importantes dados en el camino.

Durante la década de los ochenta, la importancia de la evaluación aumentó todavía más. Fue una época de gran preocupación por la calidad de la educación estadounidense. Existía una discusión generalizada sobre la evidencia desalentadora de las pruebas de logro: una disminución nacional de los resultados, en particular en el SAT, durante las décadas de los sesenta y los setenta; un desempeño en NAEP que muchos consideraban inadecuado; y comparaciones internacionales que mostraban que los resultados de logro en Estados Unidos eran inferiores a los de otros países. Esas preocupaciones condujeron a la publicación de un informe de enorme influencia –*Una nación en riesgo*– y al auge nacional de la transformación educativa conocida en ese tiempo como "el movimiento de la reforma educativa".[7] Las reformas se caracterizaban por una mayor confianza en la evaluación, el salto de las pruebas de competencia mínima a pruebas más difíciles y un cambio en las consecuencias vinculadas con los resultados de las pruebas. Aunque en algunos lugares (como Indiana y Texas) se mantuvieron las sanciones para los estudiantes individuales, la década de los ochenta presenció un cambio hacia las sanciones contra los educadores y las escuelas, como las políticas que permitían a los organismos estatales hacerse cargo de la administración de las escuelas o distritos que tuviesen un mal desempeño en las pruebas. Varios estados, como California e Indiana, experimentaron con la oferta de recompensas financieras para las escuelas con mejores resultados en las pruebas. En esa época, esas políticas se consideraban revolucionarias; 10 o 12 años más tarde se habían vuelto comunes.

A finales de la década de los ochenta nos encontramos con la primera discusión pública de calificaciones exageradas en las

pruebas de alto impacto. A John Cannell, un médico de Virginia Occidental, le desconcertaba la depresión de algunos de sus pacientes adolescentes, quienes se quejaban de tener problemas en la escuela pero que aun así tenían buenas calificaciones. Junto con su pequeño equipo, Cannel empezó a investigar y descubrió que la mayoría de los distritos y estados reportaban resultados promedio que estaban por arriba de la media nacional.[8] Este fenómeno pronto llegó a conocerse como el "Efecto del Lago Wobegon" por la mítica ciudad de Garrison Keillor donde "todas las mujeres son fuertes, todos los hombres son bien parecidos y todos los niños están por arriba del promedio".

En las publicaciones técnicas, esta exageración se denomina *inflación de las calificaciones*: incrementos en las notas que son mayores que las mejoras del logro que se supone deben señalar. La inflación de los resultados ha sido tema de un intenso debate. Muchos defensores de la evaluación de alto impacto desestiman el problema por considerarlo poco importante, pero están equivocados. Aunque la investigación sobre el tema es limitada, de manera sistemática se encuentra una considerable inflación de los resultados. Los estudios también han empezado a arrojar luz sobre los factores (distintos al simple engaño) que la ocasionan, tales como enfocar la instrucción en el material enfatizado por la prueba a expensas de otros aspectos importantes del currículo; enfocarse en detalles poco importantes de una prueba particular y enseñar trucos para presentar exámenes. En lo personal me preocupa la inflación de los resultados porque he investigado este problema por más de 15 años y porque creo que es uno de los obstáculos más serios que debemos superar si queremos encontrar maneras más eficaces de usar las pruebas para la rendición de cuentas. El capítulo 10 ofrece una revisión más detallada de la inflación de los resultados y al final del libro vuelvo a revisar algunos usos delicados de los mismos.

A finales de la década de los ochenta ocurrió otro cambio importante en los programas de evaluación de gran escala: el esfuerzo generalizado por complementar o sustituir el formato de opción múltiple por otras formas de examen, muchas de las cuales se agruparon bajo el rótulo de "evaluación del desempeño" o, más vago todavía, "evaluación auténtica". Las nuevas pruebas eran diversas y presentaban a los estudiantes muchos tipos de tareas que incluían reactivos de respuesta construida corta, reactivos que requerían respuestas escritas más extensas, desempeño práctico (por ejemplo, con algún aparato científico), tareas en que una parte del trabajo se realizaba en un grupo, y el resto era realizado por el estudiante solo, y las evaluaciones de portafolios mencionadas antes.

Algunas de las tareas individuales en esas pruebas eran de alcance modesto y podían completarse con rapidez, pero otras eran largas y complejas. Algunas requerían sesiones de varios días escolares para completarse. Por ejemplo, en 1996 se introdujeron en Nueva York diversas tareas de desempeño para la evaluación de ciencia del quinto al octavo grados. Una de esas tareas, llamada "deslizamiento", se describió de la siguiente manera:

> Los estudiantes observarán, medirán y graficarán un modelo de movimiento de pendiente descendente lenta que represente el deslizamiento del suelo.
>
> Esta tarea evalúa las habilidades del estudiante para recoger, registrar y organizar datos, disponer los ejes de la gráfica, marcar puntos de datos, dibujar gráficas lineales, aplicar las matemáticas, hacer inferencias a partir de datos de las observaciones, hacer predicciones a partir de un modelo y aplicar los modelos a otras situaciones.
>
> Esta tarea se diseñó para que los estudiantes la completaran en alrededor de 30 o 40 minutos.

Para su realización se entregaba a los alumnos un vaso de precipitado de 250 mililitros, un material viscoso con la etiqueta "mezcla", una escala métrica, un cronómetro y materiales para construir una rampa. Se les dijo que midieran el avance de la mezcla en su descenso por la rampa y que tabularan los datos obtenidos. Luego debían responder diversas preguntas.

Algunas de las preguntas eran muy limitadas y con una vinculación cercana a la tarea, por ejemplo:

> 3.a Calcula la tasa de movimiento de la mezcla durante los primeros tres minutos de observación hasta el décimo más cercano de un centímetro/minuto. Muestra tu trabajo. Tasa = distancia/tiempo.

Para encontrar esta respuesta sólo se requería la aplicación mecánica del algoritmo de cálculo proporcionado en la pregunta. Pero otras preguntas requerían conocimiento o especulación que iba mucho más allá de la propia actividad, por ejemplo:

> 5.b Esta actividad presentó un modelo para movimientos descendentes como los aludes, deslizamiento del suelo o actividad glacial. En la naturaleza, ¿qué podría suceder para incrementar la tasa de movimiento del sedimento o del hielo en esas características de la tierra?*

---

* Este pasaje, que es todo lo que el espacio permite, no hace justicia a la complejidad y diversidad de las tareas de evaluación del desempeño. Para apreciarlas en toda su esencia, deben buscarse en las direcciones administrativas las descripciones completas (a menudo con ilustraciones) de las tareas y las explicaciones de las rúbricas usadas para calificar el desempeño del estudiante. Para encontrar una colección fascinante de tareas de evaluación del desempeño en ciencia, vaya al sitio de los Vínculos de Evaluación del Desempeño en Ciencia mantenido por SRI International en *http://pals.sri.com/*. La tarea descrita puede encontrarse en *http://pals.sri.com/tasks/5-8/Creeping/* (consultada por última vez el 8 de julio de 2006).

Los defensores de la evaluación del desempeño en ocasiones se dejaban llevar por el entusiasmo. Por ejemplo, algunos sostenían que en condiciones ideales las pruebas debían elaborarse a partir de tareas que no tuviesen una única respuesta correcta, porque los problemas importantes que enfrentan los adultos en la vida real no tienen respuestas correctas únicas. En un congreso de expertos en evaluación realizado en esos años, una destacada defensora de la evaluación del desempeño ofreció una conferencia en que presentaba justamente este argumento. En una forma tan extrema, es una posición tonta; si bien es cierto que muchos problemas reales no tienen respuestas correctas únicas, muchísimos otros sí las tienen. Por ejemplo, el piloto que llevó a la conferencista al congreso tuvo que decidir si durante el aterrizaje los alerones debían estar arriba o abajo. Por fortuna para la conferencista, el piloto eligió la única respuesta correcta. Cuando la oradora salía del edificio después de su presentación, se detuvo y me dijo que no sabía cuál de los hoteles cercanos era el Hilton, donde tenía una reservación. Señalé que esa era una pregunta con una única respuesta correcta, ante lo cual se fue enojada sin permitir que le dijera cuál era el hotel.

El entusiasmo por la evaluación del desempeño, que floreció con notable rapidez, tuvo varias raíces. Una estaba implicada en el término autenticidad, un deseo de evaluar el desempeño de los estudiantes en tareas realistas de complejidad similar a las que pueden encontrar fuera de la escuela. Esto fue una reacción al diseño de las pruebas tradicionales en que las habilidades y el conocimiento se descomponen de manera deliberada en pequeñas piezas. Por lo general se presentaba a la evaluación auténtica como algo nuevo, incluso revolucionario, pero la evaluación del desempeño era cualquier cosa menos nueva y muchos de los argumentos a favor o en contra de esta aproximación a la evaluación habían sido expuestos claramente casi medio siglo antes en el artículo de E. F. Lindquist descrito en el capítulo anterior.

Una segunda razón del interés que hubo en esos años por la evaluación del desempeño fue el énfasis entre los reformadores de la educación en las metas relacionadas de establecer estándares elevados para todos los estudiantes y enfocar la enseñanza en habilidades de orden superior más que en habilidades básicas y conocimiento factual; por ejemplo, en matemáticas se enfatizaba la solución de problemas, el razonamiento y la comunicación en lugar de la simple aplicación de procedimientos aritméticos. En ese tiempo existía la idea generalizada de que las evaluaciones del desempeño eran más adecuadas que las pruebas de opción múltiple para medir esas habilidades superiores. Hay algo cierto en esta opinión, pero es demasiado simple. La investigación ha demostrado que el formato en que se presentan las tareas a los estudiantes no siempre predice de manera confiable qué habilidades utilizarán y es común que no logren aplicar habilidades de orden superior a la solución de tareas que parecen requerirlas.

Una tercera razón del atractivo de la evaluación del desempeño fue la creencia cada vez más común de que "lo que se examina es lo que se enseña". Los defensores sostuvieron que cuando se hace a la gente responsable de los resultados en las pruebas de opción múltiple, se fomenta una enseñanza que se asemeje a esas pruebas –lectura de pasajes breves, problemas cortos, preguntas de opción múltiple en lugar de otras que requieran escritura, etcétera– y que este tipo de enseñanza es aburrida, poco exigente cognitivamente e improductiva. Es indudable que esto sucedía en cierta medida, no como resultado de la evaluación de opción múltiple como tal, sino como consecuencia de hacer que los profesores se preocuparan demasiado por los resultados obtenidos en esas pruebas. En ese tiempo un reportero dedicó varias semanas a observar escuelas de los suburbios de Washington, D.C. que recibían calificaciones inusualmente altas y que atendían sobre todo a estudiantes de grupos minoritarios con bajos ingresos, más tarde me dijo que los

maestros de esas escuelas empezaban a preparar a los estudiantes para la prueba de opción múltiple de tercer grado desde que estaban en el jardín de niños.

Los defensores de la evaluación del desempeño argumentaban que la respuesta a esta situación debería ser que "las pruebas merecieran que la enseñanza se enfocara en ellas". Deberían ser pruebas que evaluaran habilidades de orden superior, como la solución de problemas complejos, en lugar de la simple aplicación de cálculos aritméticos; y deberían insertar esas habilidades en tareas realistas y complejas. Los reformadores arguyeron que las nuevas pruebas alentarían la instrucción no sólo al evaluar contenidos abundantes y exigentes, sino también al modelar tipos de tareas que contribuyen a la buena enseñanza. Es decir, las mismas tareas de la evaluación ejemplificarían los tipos de trabajo que los maestros deben incluir en su enseñanza. Esto representaba un cambio importante en la noción subyacente de la manera en que las pruebas deben ayudar a mejorar la enseñanza. En la evaluación del logro tradicional las tareas se diseñaban para obtener información diagnóstica que permitiera a los maestros mejorar su práctica docente, pero no se esperaba que las tareas usadas en la enseñanza fuesen parecidas a las de la prueba.

La frase "que las pruebas merezcan que la enseñanza se enfoque en ellas" tenía también otra connotación: debían diseñarse de modo que la preparación de los estudiantes para presentarlas –enseñar para la prueba– no diera lugar a la inflación de los resultados. Pero esto era una prestidigitación lógica. No hay razón para esperar que una prueba que "vale la pena que se enseñe para ella" en el sentido de medir habilidades de orden superior y cosas similares fuese inmune a la inflación de resultados. Además, como explico en el capítulo 10, la investigación ha confirmado que esta expectativa era falsa: puede darse una considerable inflación incluso en los resultados obtenidos en pruebas que evitan el formato de opción múltiple.

Aunque el movimiento de evaluación del desempeño ha tenido efectos duraderos en los programas de evaluación a gran escala, los políticos muy pronto se alejaron de las formas más extremas de evaluación del desempeño. Además de ser costosa y llevarse mucho tiempo (los estudiantes tardan mucho más tiempo para completar las tareas complejas), este tipo de evaluación plantea graves dificultades técnicas. Por ejemplo, a menudo es difícil calificar de manera confiable las evaluaciones del desempeño y es difícil (en algunos casos impráctico) hacer que los resultados tengan un significado comparable de un año a otro o de una escuela a otra.

Otro cambio en la forma de evaluación –a primera vista, algo que sólo podría gustarle a un psicómetra, pero que en realidad es muy importante y en ocasiones controvertido– fue la generalización de las *evaluaciones por muestreo matricial*. En la evaluación estandarizada convencional, a todos los estudiantes de un determinado tipo (por ejemplo, todos los alumnos en un salón regular de quinto grado) se les aplicaban precisamente los mismos reactivos de examen. En la evaluación por muestreo matricial, la prueba se descompone en distintas partes que incluyen diferentes tareas, las cuales se distribuyen luego al azar en las aulas o escuelas. De modo que la prueba no se estandariza para comparar a estudiantes individuales, sino que se estandariza con el propósito de comparar escuelas o estados. El muestro matricial ahora es común; por ejemplo, se utiliza en NAEP y TIMSS, y algunas evaluaciones estatales.

La importancia de esta innovación aparentemente misteriosa es que permite la evaluación de una mayor variedad de conocimientos y habilidades (una muestra mayor del dominio) en un determinado tiempo de evaluación. Al inicio, este enfoque se consideraba ventajoso por el solo hecho de que ofrecía mayor información, pero cuando las pruebas se usan para fines de la rendición de cuentas ofrece otro beneficio de trascendental importancia: cambia los incentivos para los estudiantes y los maestros. Entre

mayor sea la amplitud de la prueba, menos incentivos hay en reducir inapropiadamente la instrucción como una forma de "ganarle" al sistema e inflar los resultados.

Muchos dirían entonces que si este cambio esotérico en el diseño de la prueba nos ofrece información más completa sobre el logro y mejores incentivos para los maestros y los alumnos, ¿por qué no permitir que los psicómetras hagan lo que saben hacer y usen de manera rutinaria el muestreo matricial? Como suele suceder en la evaluación, hay un precio que pagar y un difícil compromiso entre metas. Un diseño puro de muestreo matricial no proporciona calificaciones útiles para los estudiantes individuales porque éstos presentan subpruebas diferentes y, por ende, no comparables. De ahí la controversia. Muchos padres han argumentado que si sus hijos van a invertir tiempo y esfuerzo para participar en una evaluación (para no hablar del tiempo mucho mayor que dedican a prepararse para la prueba), por lo menos quieren una calificación a cambio. La prueba del Sistema de Evaluación Integral de Massachusetts (*Massachusetts Comprehensive Assessment System*) ejemplifica un compromiso: una parte de la prueba es común para todos los estudiantes y se utiliza para proporcionar calificaciones o resultados individuales, mientras que el resto es una muestra matricial y sólo contribuye a los resultados de las escuelas.

Una oleada más duradera de la reforma educativa que coincidió con la era de la evaluación del desempeño fue el cambio a las pruebas *basadas en estándares* o *referidas a estándares*. La idea que subyace a esta forma de evaluación es que los estados o localidades deben empezar por especificar los *estándares de contenido* (que son planteamientos de lo que los estudiantes deben conocer y poder hacer) y los *estándares del desempeño*, que son declaraciones de qué tan bueno se espera que sea el desempeño de los estudiantes con respecto a los estándares de contenido. Las pruebas deben

entonces alinearse con los estándares de contenido y diseñarse para determinar qué estudiantes alcanzaron los estándares de desempeño.

Los defensores de la evaluación referida a estándares consideran que es un cambio importante respecto de la evaluación tradicional. (Una señal es que algunos evitan el término prueba y en lugar de ello se refieren a sus medidas como "evaluaciones".) Pero en realidad, las pruebas basadas en estándares se alejan menos de la tradición de lo que muchos creen. La elaboración de las pruebas tradicionales de logro de alta calidad también empieza con grandes esfuerzos por aclarar lo que los estudiantes deben saber y poder hacer, aunque a eso se le suele llamar "marco curricular" en lugar de "estándares de contenido". En la práctica, el contenido y el formato de las pruebas estatales basadas en estándares varían de manera significativa; algunas son muy parecidas a las pruebas convencionales y en la elaboración de la mayor parte de ellas se emplean métodos similares (por ejemplo, procedimientos estadísticos semejantes).

Sin embargo, los estándares de contenido tienen una diferencia importante respecto de los marcos curriculares de las pruebas tradicionales: por lo general son específicos a estados individuales. Las pruebas de logro tradicionales se diseñaron buscando elementos comunes en los marcos curriculares de muchos estados, de modo que los editores pudieran comercializar de manera amplia sus pruebas y proporcionar normas nacionales. Los defensores de la evaluación basada en estándares por lo regular exigen que las pruebas estén alineadas con los estándares particulares de cada estado ya que –según su argumento– el mayor alineamiento producirá mayor claridad acerca de las metas educativas de los estados y aumentará la sensibilidad de las pruebas a las mejoras en la educación. Esto genera presión para que los estados utilicen pruebas distintas. Sin embargo, no existe nada gratis y este beneficio tiene

dos costos importantes: empeora el problema del cruce del Potomac con el que inicié este capítulo, e incrementa considerablemente el volumen de pruebas que deben elaborarse, lo que ejerce mucha presión sobre la capacidad de la pequeña industria de la evaluación y crea el riesgo de pruebas de menor calidad.

Al apoyarse en las pruebas precedentes de competencia mínima, la evaluación basada en estándares también se aparta de la tradición en la manera en que se reporta el desempeño. En esas pruebas el desempeño se reporta sobre todo en términos de si los estudiantes alcanzaron uno o más de los estándares de desempeño. En un sistema típico basado en estándares, se ubica a los educandos en una de cuatro categorías: no logró alcanzar el estándar más bajo, superó el más bajo pero no cumplió con el segundo, aprobó el segundo (por lo regular, "competente"), pero no llegó al tercero, y superó el más alto. En contraste, el reporte tradicional depende de diversas escalas numéricas que hacen posible un gran número de resultados.

Esta innovación en la calificación ahora se acepta de manera casi universal, se incorporó en la legislación federal y en general se considera deseable porque se enfoca en las expectativas y supuestamente es sencilla de entender. Sin embargo, su costo es demasiado alto, tal vez excesivo. El proceso de establecimiento de estándares (decidir cuánto tienen que hacer los estudiantes para estar a la altura del estándar) es técnicamente complejo y tiene un aura científica, pero los estándares son de hecho muy arbitrarios. La sencillez de esta forma de reporte es, por ende, más aparente que real y la mayoría de la gente no tiene en verdad una idea clara de lo que en realidad significan los estándares. Por esta razón, es común encontrar el regreso sigiloso al reporte referido a normas, por ejemplo, para mostrar cómo se compara el porcentaje de "competencia" de una escuela con el de otros planteles. El informe basado en estándares proporciona una visión muy burda y, en

algunos casos, muy distorsionada del logro y puede crear el indeseable incentivo de dedicar más atención a los niños que se acercan más al estándar que cuenta, en detrimento de los otros niños (esos problemas se analizan más a fondo en el capítulo 8). Puede ser útil saber si los estudiantes dan la medida, pero confiar demasiado en los estándares de desempeño (en particular, usarlos por sí solos sin otras formas más tradicionales de reporte) es una receta para problemas.

Al inicio de la década de los noventa se observó un rápido incremento en los esfuerzos por incluir a más estudiantes con necesidades especiales (alumnos con discapacidades y con una competencia limitada del inglés) en las evaluaciones aplicadas por los estados a sus alumnos de educación general.* De manera tradicional, muchos de estos estudiantes eran excluidos de las evaluaciones porque no se consideraba que éstas guardaran relación con su programa educativo, porque las pruebas eran demasiado difíciles para ellos, debido a que sus necesidades especiales hacían que la forma estándar de la prueba fuese inapropiada para ellos o al simple hecho de que su inclusión derribaría los resultados promedio. La lógica para incrementar su participación fue sencilla: si las pruebas se utilizan para responsabilizar a los educadores de mejorar el logro de sus alumnos, los maestros tendrían poco incentivo para prestar atención al logro de los estudiantes con necesidades especiales, a menos que también se les evaluara. Este cambio se instituyó primero con un alcance estatal, por ejemplo,

* El término *competencia limitada del inglés* (CLI) ha caído en desuso, ahora suele hablarse de *aprendiz del inglés* (AI). Sin embargo, como explico en el capítulo 12, el problema para los propósitos de la buena evaluación no es el hecho de que un estudiante esté aprendiendo inglés; eso es algo que hacen de pequeños incluso los hablantes nativos. Lo que es importante para la evaluación es si los hablantes no nativos han alcanzado un nivel de competencia que les permita hacer una demostración adecuada de su conocimiento en una prueba.

en Kentucky y Maryland. Con el tiempo se convirtió en tema del estatuto federal con las Enmiendas hechas en 1997 a la Ley de Educación para los Individuos con Discapacidades y la ley de 2001 «Que ningún niño se quede atrás».[9]

Al terminar el siglo XX, la evaluación estandarizada del logro exigida por el estado era casi universal en Estados Unidos. La mayoría de los estados habían creado sus propios estándares de contenido y de desempeño, aunque muchos usaban pruebas comercializadas nacionalmente (algunas veces con modificaciones) que consideraban lo bastante alineadas con sus estándares en lugar de usar pruebas totalmente adaptadas. La mayor parte de los programas reportaban el logro mediante la comparación con los estándares de desempeño, aunque muchos usaban también otras escalas de reporte más convencionales. La mezcla de formatos variaba, pero era común encontrar una combinación de preguntas de opción múltiple y tareas modestas de respuesta construida, como preguntas de respuesta corta y ensayos. Casi todos los estados usaban resultados para recompensar y castigar a las escuelas de alguna manera, y alrededor de la mitad (la cuenta cambiaba de manera casi continua) usaba al menos una prueba de alto impacto para los estudiantes, por lo regular como requisito para la graduación de la preparatoria.

Los estados establecieron diversos métodos para hacer a los educadores responsables de los resultados. Algunos establecieron sencillamente un estándar de desempeño y luego monitoreaban qué escuelas lo cumplían. Otro método, raro pero que atraía cada vez más interés, fue un enfoque de *valor agregado* en que se sigue la trayectoria escolar de los estudiantes y se evalúa a las escuelas o a los maestros en relación con de los progresos de sus alumnos de un grado al siguiente. Este método es intuitivamente atractivo, pero enfrenta muchos obstáculos desalentadores en términos tanto de la evaluación como de la maquinaria estadística usada para analizar los resultados.

No obstante, el método más común empleado por los estados durante la década de los noventa consistió simplemente en comparar el desempeño de los estudiantes en un determinado grado con cohortes anteriores de alumnos del mismo grado. Por ejemplo, se compararía el porcentaje de alumnos de cuarto grado que este año alcanzaron el estándar estatal de competencia con el porcentaje de alumnos de cuarto grado que obtuvieron ese nivel el año pasado. Este enfoque tiene muchas ventajas, entre ellas la sencillez, pero también varios defectos. Uno es que los resultados de cualquier cohorte de estudiantes no sólo son determinados por la calidad de su educación sino también (y de manera poderosa) por factores no educativos como los antecedentes sociales. Las escuelas que atienden a estudiantes con carencias obtendrán resultados más bajos que escuelas de eficacia comparable que atienden a niños más favorecidos. Los efectos de los cambios en las características de una comunidad, como una inmigración rápida y otras tendencias demográficas, se confunden con los cambios en la eficacia educativa.

Si se va a hacer responsables de los resultados a los educadores, alguien tiene que decidir cuánta mejora es suficiente. Estos objetivos por lo general se eligieron sin contar con todos los elementos, sin base en evidencia sólida como datos normativos, comparaciones internacionales, tendencias históricas que muestren la probabilidad de que se den mejoras rápidas a lo largo del tiempo o evaluaciones de intervenciones a gran escala. En algunos casos, las expectativas resultantes simplemente no eran razonables. Por ejemplo, muchos reformadores sostenían que debía esperarse que los estudiantes alcanzaran un nivel de desempeño similar al nivel de competencia establecido por la Evaluación Nacional del Progreso Educativo. Pero en Estados Unidos, menos de uno de cada cuatro estudiantes alcanza el nivel competente en NAEP de matemáticas de octavo grado, y el desempeño en TIMSS sugiere que en

números redondos, alrededor de 30 por ciento de los estudiantes de Japón y Corea también quedarían muy por debajo en la misma evaluación.[10] ¿En verdad esperamos que en el corto plazo o incluso en un plazo moderado, sea posible elevar el desempeño de todos los estudiantes a un nivel que en la actualidad no logran alcanzar más de tres cuartas partes de los estudiantes de Estados Unidos y alrededor de un tercio de los estudiantes de dos de los países con mayor logro del mundo? Encontré que en un estado se esperaba que la escuela típica lograra mejoras que en dos décadas habrían puesto a más de la mitad de los alumnos por arriba de un nivel que al inicio sólo alcanzaba 2 por ciento de los estudiantes y se esperaba que el progreso de las escuelas con bajos resultados fuese todavía mayor. También puede apreciarse el enorme tamaño de la ganancia que se espera de la escuela promedio comparando algunas de las diferencias grupales particularmente grandes en el desempeño que se aprecian en los datos actuales. La mejora esperada era dos veces mayor que la diferencia promedio en matemáticas entre Estados Unidos y Japón en TIMSS. También era el doble de la diferencia promedio que suele encontrarse entre los estudiantes afroamericanos y los estudiantes blancos no hispanos. Ninguna investigación sugiere que pueda tenerse confianza de lograr ganancias de esta magnitud a gran escala.

En un sistema como este, también se necesita decidir la rapidez y constancia con que las escuelas deben progresar hacia el objetivo. Muchos estados establecen esas expectativas usando el método de "línea recta". Con la meta de que todos los estudiantes alcanzarían el estándar de competencia al final de un determinado periodo, se establecieron los objetivos intermedios trazando una línea recta entre el porcentaje inicial de competencia y el 100 por ciento de competencia al final del tiempo señalado. Esto requería tasas arbitrarias de mejora y asumía (sin evidencia) que las escuelas que al inicio tenían un bajo desempeño podían mantener tasas de

mejora mucho más rápidas que las escuelas con altos resultados. También se suponía que los maestros tienen la habilidad de crear mejoras sistemáticas e ininterrumpidas. Como alguien que ha enseñado en casi todos los niveles, desde cuarto grado hasta los estudios de doctorado, por lo general con evaluaciones muy favorables, me parece que esta última expectativa es notablemente poco realista. Las mejoras reales en la instrucción suelen ser irregulares. Si un maestro reconoce que una técnica para enseñar un tema no ha funcionado bien, no existe garantía de que la primera alternativa que utilice será adecuada, e incluso si lo es, tal vez necesite probarla en varias ocasiones para hacerla funcionar. Es difícil experimentar con nuevos métodos que son prometedores pero que implican riesgos si se espera que uno logre un aumento constante en los resultados.

Además, esos métodos para mejorar el desempeño impusieron expectativas idénticas para todas las escuelas que tenían los mismos resultados iniciales. No se hizo ningún esfuerzo para determinar los factores particulares que daban lugar a un bajo desempeño en una escuela determinada o para adaptar las expectativas de mejoras a condiciones específicas. Por ejemplo, considere dos escuelas hipotéticas con resultados similares e inaceptablemente bajos. La escuela A tiene una población estudiantil estable compuesta por angloparlantes nativos pero un cuerpo docente muy malo. La escuela B tiene un mejor profesorado, pero a esto se le opone el hecho de que una gran proporción de sus alumnos son inmigrantes que, además de no ser competentes en inglés, provienen de muchos antecedentes lingüísticos, lo que hace imposible encontrar maestros bilingües para todos ellos. En la mayoría de los estados, los objetivos de mejora para esas escuelas habrían sido idénticos.

Lo extremo de las expectativas poco realistas de este enfoque se hace más claro si uno imagina tratar de hacer algo similar en otra área de interés público, como la calidad de los hospitales. Primero

estableceríamos estándares de "resultados de salud suficiente", usando métodos arbitrarios y diferentes en los distintos estados que arrojaran respuestas diversas que no se basaran en ninguna evidencia sobre lo que la tecnología médica actual podría producir. Luego diríamos a todos los hospitales, sin tener en cuenta sus circunstancias (por ejemplo, la edad o condición de salud de sus pacientes, la cantidad de especialistas disponibles en su área geográfica, los recursos a los que tienen acceso, etcétera) que tienen un tiempo establecido, digamos 12 años, para alcanzar el punto en que todos los pacientes sean dados de alta en un estado de "salud suficiente". En el trayecto, serían recompensados o castigados según los progresos lineales que hacen hacia esta meta. Es difícil imaginar que semejante propuesta fuese considerada con seriedad.

Esto nos lleva a la ley «Que ningún niño se quede atrás», que es la más reciente de las nuevas autorizaciones periódicas de la Ley de Educación Primaria y Secundaria de 1965. La propuesta para dicha ley provino de la Casa Blanca, que es presentada por la administración y en general así considerada como iniciativa del presidente Bush. Sin embargo, su historial político es un poco más complejo. Si bien dicha ley fue ciertamente iniciativa del presidente Bush, se promulgó con apoyo de ambos partidos, en cierta medida por el respaldo de dos influyentes miembros liberales del Congreso: el senador Edward Kennedy de Massachusetts y el diputado George Miller de California. Según me contó Miller algún tiempo después de la aprobación del proyecto de ley, durante décadas habían fracasado los esfuerzos para mejorar el logro de los estudiantes desfavorecidos y era tiempo de "arrojar algo de luz en las esquinas."

En cierto sentido es una ley innovadora, pero en otros representa una continuación de las tendencias evaluativas que la precedieron. La NCLBA combinó elementos de políticas estatales comunes para la evaluación y la rendición de cuentas, agregó algunos otros e

hizo del paquete una obligación federal para cualquier estado que quisiera recibir financiamiento del Título I de la Ley de Educación Primaria y Secundaria. (Según la interpretación común de la Constitución, el gobierno federal tiene muy poco poder para exigir prácticas o políticas educativas, por lo que el mecanismo para obligar a los estados a cumplir las previsiones de la ley NCLB consistió en condicionar los importantes fondos del Título I a su cumplimiento. En respuesta, varios estados consideraron renunciar a los fondos del Título I, aunque ninguno lo ha hecho todavía.) La mencionada ley requiere que entre los grados tercero y octavo se hagan evaluaciones anuales en matemáticas y lectura y en un grado de secundaria. Pronto entrará en efecto el requisito de hacer evaluaciones de ciencia en un mínimo de tres grados. Se pretende que todos los estados usen para este propósito una prueba referida a estándares y que los resultados se reporten en términos de estándares de desempeño, incluyendo uno llamado competente. Exige que al cabo de 12 años el desempeño de prácticamente todos los estudiantes alcance el nivel competente y establece un sistema complejo para determinar si los estados y las escuelas están haciendo un "progreso anual adecuado" ("*adequate yearly progress*"-AYP) hacia la meta. (El AYP al parecer se basa en los modelos de crecimiento de línea recta de algunos estados, aunque difiere en el hecho de que exige el establecimiento de un solo objetivo estatal cada año para todas las escuelas del estado.)

La NCLBA exige que casi todos los estudiantes, incluso la mayoría de los que tienen necesidades especiales, sean evaluados con la misma prueba y que casi todos sean sometidos a los mismos estándares de desempeño. Requiere que se hagan reportes separados a niveles del estado y de la escuela para grupos raciales y étnicos, para estudiantes con competencia limitada del inglés, para estudiantes con discapacidades y para estudiantes con carencias económicas, a menos que el número de alumnos en un grupo sea

tan pequeño que los resultados serían poco confiables y considera que una escuela no está logrando un AYP si alguno de esos grupos no logra hacerlo. A eso se refería el diputado Miller cuando hablaba de arrojar luz en las esquinas. La ley también exige que se apliquen sanciones a las escuelas que no logren hacer un AYP y especifica un conjunto de sanciones cada vez más severas si dicho fracaso persiste.

Es difícil exagerar el impacto de la NCLBA en la educación primaria y secundaria durante su corta vida. En muchos sentidos, es la culminación de la transformación que empezó por lo menos con el movimiento de evaluación de competencia mínima y quizá antes, del uso de las pruebas de logro para fines de diagnóstico y evaluación local al uso de las pruebas como un medio de evaluar sistemas educativos enteros y de hacer a los maestros responsables de los cambios en los resultados de las pruebas. Independientemente de los argumentos a favor y en contra de la responsabilidad basada en las pruebas, creo que permanecerá entre nosotros en el futuro inmediato. Por lo tanto, al final del libro, cuando regrese a los usos razonables de las pruebas, una de las cuestiones primordiales será la mejor manera de usar las pruebas para hacer a los maestros responsables. ■

El ABC
de la
evaluación educativa

Capítulo 5
¿Qué nos dicen las calificaciones de las pruebas sobre los niños estadounidenses?

Durante más de un cuarto de siglo el debate sobre la educación estadounidense ha estado dominado por los resultados obtenidos en pruebas estandarizadas. Según a quien se escuche, los resultados se han utilizado para decirnos que el logro de los estudiantes de este país ha disminuido; que está o que no está repuntando; que la brecha entre los estudiantes de los grupos minoritarios y mayoritarios se está o no estrechando, y que el desempeño de sus alumnos es o no es lo bastante bueno en comparación con los alumnos de otros países. Los resultados obtenidos en las pruebas han sido el foco de atención de varios informes destacados sobre educación y reformas económicas. Es frecuente que los periódicos pongan en primera plana los resultados de los programas estatales y locales de evaluación y den cobertura amplia a las mediciones del logro nacionales e internacionales.

Sin embargo, puede entenderse la confusión de cualquiera que trate de seguir esta información en las versiones periodísticas, los comunicados de prensa o las declaraciones públicas de los reformadores de la educación o de los administradores distritales y estatales. Los informes suelen ser incongruentes, incluso cuando se hace referencia a los mismos datos. Las afirmaciones sobre los resultados a menudo son simplemente erróneas. Es común que la manera de reportar los resultados haga difícil saber si un cambio en los mismos o una diferencia entre grupos constituyen noticias

relativamente buenas o inusualmente malas. Por lo general se ignoran por completo los cambios en el contexto que debería determinar la interpretación de los resultados (como las tendencias en la mezcla de estudiantes evaluados), a la vez que se difunden afirmaciones sin fundamento acerca de las causas de los cambios en el desempeño.

Es importante conocer bien los hechos. En este capítulo voy a recurrir a los datos de los pasados 40 años para describir las tendencias en el logro de los estudiantes estadounidenses y para explorar cómo se comparan con los de otros países. En el siguiente capítulo analizaré los factores que influyen en los resultados obtenidos en las pruebas, en particular, la idea común de que la educación tiene un efecto tan grande en los resultados que puede suponerse que las escuelas con mejores resultados ofrecen una mejor educación. Pero primero es necesario explicar una métrica de uso común para evaluar el tamaño de las diferencias en los resultados y para comparar los hallazgos de una prueba a otra.

Una escala común para pruebas diferentes

En las mediciones que encontramos en casi todos los aspectos de la vida diaria usamos escalas que son tan familiares, como las pulgadas o los centímetros, que pensamos poco en ellas. Si escuchamos que un hombre es media pulgada más alto que otro, sabemos que sus estaturas son muy similares porque entendemos la escala de pulgadas para longitud. Si nos enteramos de que la temperatura máxima del día será de 95 grados Fahrenheit en lugar de 72, sabemos que en la tarde hará un calor sofocante en lugar de ser deliciosamente cálida. Sólo nos desconcertamos cuando nos encontramos con una escala desconocida, por ejemplo, cuando los viajeros estadounidenses se topan con la escala de temperatura

que se usa prácticamente en todo el mundo civilizado que les obliga a entender que una temperatura de 35 grados Celsius es demasiado caliente, o cuando los turistas que llegan a Estados Unidos enfrentan las temperaturas veraniegas de 95 grados, que consideran cercanas al punto de ebullición del agua.

Las escalas usadas para reportar el desempeño de los estudiantes en las pruebas no son como estas. Diferentes pruebas se reportan en escalas distintas y muchas de ellas son arbitrarias. Considere las dos pruebas rivales de admisión a la universidad: la SAT y la ACT. La escala de matemáticas de la prueba SAT va de 200 a 800, mientras que la escala correspondiente de la prueba ACT va de 1 a 36. ¿Qué indica esta diferencia en las escalas? Nada en absoluto. Esas escalas son arbitrarias, no tienen significado intrínseco y no son comparables. Para comparar una puntuación de 25 de la ACT con una de 700 en la SAT es necesario convertir una o ambas calificaciones para ponerlas en una escala común.

Además de ser arbitrarias, las escalas que miden los resultados de la mayoría de las pruebas no son lo bastante conocidas para que su significado resulte intuitivamente claro. Por ejemplo, entre 1999 y 2004, el resultado promedio en lectura de los niños de nueve años en la evaluación de la tendencia a largo plazo mediante la Evaluación Nacional del Progreso Educativo (NAEP) aumentó 7 puntos, de 212 a 219. ¿Eso es mucho o es poco? Aunque esta escala en particular ha aparecido muchas veces en las noticias durante más de dos décadas, pocos saben si un incremento de 7 puntos es importante. Ni siquiera las escalas de resultados más conocidas resultan del todo claras para la mayoría de la gente. Al menos en los estados en que la prueba SAT es la que más se utiliza para admisión a la universidad, pocas escalas de resultados son más conocidas que la de dicha prueba. Los padres de estudiantes que solicitan ser admitidos a universidades selectivas en esos estados saben que un resultado en matemáticas de 750 en la SAT-I es

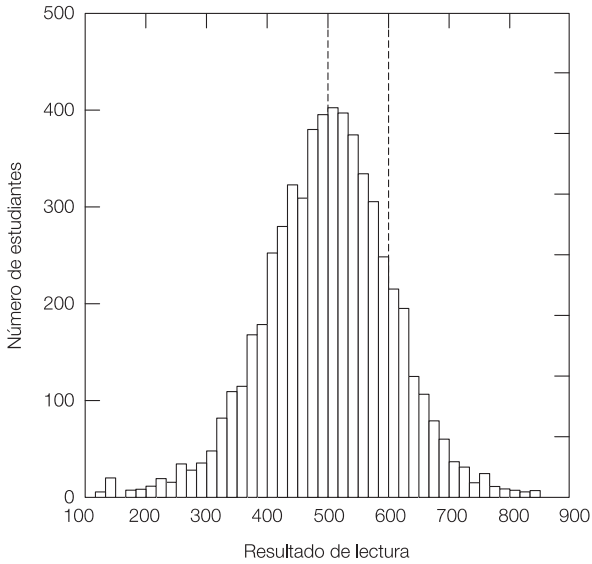
muy alto. Pero si el resultado promedio de la prueba SAT se incrementa o disminuye en 35 puntos ¿es eso un cambio considerable? Comparar los resultados de diferentes pruebas es todavía más difícil. Si el resultado promedio de un grupo en la sección de matemáticas de la prueba ACT aumenta en dos puntos de 36 posibles, ¿cómo se compara esto con un incremento de 22 puntos en la prueba SAT, que tiene nivel máximo de 800? Es imposible responder esas preguntas a menos que esos números se conviertan en otra forma, aunque eso no impide que muchas personas reporten cambios en esas escalas como si su significado fuese claro.

La solución más común a este problema en la evaluación educativa, como en muchas otras ciencias, es convertirlo todo a una única escala, bien conocida, que se basa en la *desviación estándar*, que es una medida de qué tanto varían o se dispersan los resultados (o cualesquier otro rasgo). Una vez que esto se ha hecho, es sencillo comparar los resultados de diferentes pruebas y también pueden traducirse a otras formas intuitivamente comprensibles, como los rangos percentiles.

Para ilustrar lo anterior, comencemos con algunos datos reales del programa de evaluación de un estado no identificado. La figura 5.1 muestra la distribución de resultados en la prueba de lectura de una escuela secundaria. Es un histograma en que la altura de cada barra representa el número real de estudiantes que reciben un resultado en ese rango particular. De modo que la barra más alta nos muestra que más de 400 estudiantes recibieron una puntuación de 500 o un poco mayor. La escala de esta prueba es arbitraria, justo como las escalas del SAT y la NAEP, y la modifiqué a partir de una que se usa en la actualidad para ocultar al estado que me proporcionó los datos.

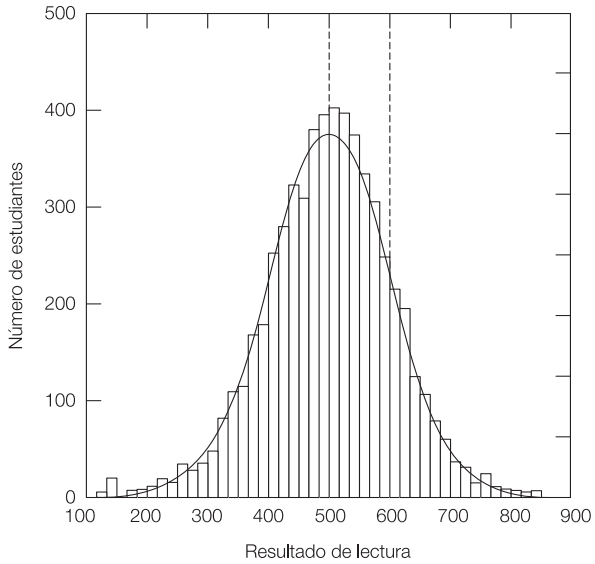
Suponga que encontramos una diferencia en esta escala (digamos, la diferencia entre las puntuaciones de 500 y 600, señalada por las dos líneas verticales punteadas de la figura 5.1) y tenemos que de-

Figura 5.1. La distribución de los resultados de lectura en un estado no identificado



cidir qué tan grande es. Sin la figura no se habría tenido idea; una diferencia de 100 puntos puede ser enorme o pequeña, dependiendo de la escala (arbitraria). Sin embargo, la gráfica nos da un indicio que no tendríamos con los números por sí solos, ya que nos permite comparar nuestra diferencia de 100 puntos con la distribución de resultados. Nos muestra que una puntuación de 500 es más o menos el promedio, mientras que una de 600 es bastante alta en relación a la distribución de resultados; es decir, pocos alumnos obtuvieron puntuaciones por arriba de 600. De igual manera, no fueron muchos los que obtuvieron resultados 100 puntos por debajo del promedio, es decir, por debajo de 400. Desplazarse de 500 a 600 implica sobrepasar a un número considerable de estudiantes, por lo que, en ese sentido, una diferencia de 100 puntos es grande. (No hay nada mágico acerca de los 100 puntos en este ejemplo. Lo elegí sólo porque es un número conveniente dada la aritmética que

Figura 5.2. La distribución de los resultados de lectura en un estado no identificado con la curva normal sobrepuesta



utilicé para calcular esta escala). El número ilustra un punto general: conocer la *distribución* de resultados proporciona información útil acerca del tamaño de cualesquier diferencia dada en los resultados.

Lo que se necesita es una manera de ser más precisos al usar la información sobre la distribución de resultados. Vamos a sobreponer a esta distribución de los resultados reales la famosa "curva de campana", mejor conocida como distribución normal o Gaussiana. En la figura 5.2 puede verse que la curva de campana se ajusta muy bien a esta distribución. Hay algunas protuberancias y bamboleos en el histograma, pero en general, la distribución no se aparta mucho de la distribución normal.

A lo largo de los años, la curva de campana ha adquirido todo tipo de connotaciones terribles. Algunos insisten en que es la creación malintencionada de los psicómetras que quieren crear la apariencia de diferencias entre los grupos. Otros la asocian con

la visión perniciosa e infundada de que las diferencias en los resultados entre grupos raciales y étnicos están biológicamente determinadas. Ninguna de esas asociaciones está justificada. La curva de campana es sencillamente una manera de describir una distribución que es muy común en la naturaleza (por ejemplo, la distribución de las circunferencias de las cabezas de las momias egipcias se ajusta aproximadamente a la curva de campana) y que tiene algunas propiedades matemáticas que resultan ser de gran utilidad.

El hecho es que en realidad no sabemos cómo debería lucir la "verdadera" distribución del logro en lectura o en matemáticas y podemos diseñar pruebas para cambiar la forma de la distribución. Por ejemplo, al modificar la forma en que se elabora o se escala la prueba podemos estirar una de las "colas" de la distribución, dando a los estudiantes con puntuaciones muy bajas o muy altas posiciones más alejadas de la media. Sin embargo, cuando se cumplen varias condiciones comunes (cuando las pruebas evalúan dominios amplios, están constituidas por reactivos que tienen un rango de dificultad razonable y se escalaron por medio de los métodos más comunes en la actualidad), los resultados de la escala suelen mostrar una distribución aproximadamente normal, con muchos alumnos que se agrupan cerca del promedio y que van disminuyendo progresivamente a medida que uno se acerca a los extremos. No obstante, no es raro encontrar excepciones. Si una prueba es sencilla para los estudiantes que la presentan, la distribución no será normal sino asimétrica, con una cola de bajos resultados pero muchos estudiantes agrupados cerca de los resultados más altos. En una prueba que sea demasiado difícil para los sustentantes los resultados estarán sesgados en la otra dirección, los resultados se concentrarán en el extremo inferior mientras una cola delgada se extiende a los resultados más altos. Por ejemplo, en un estudio de un programa de evaluación estatal que realicé en la década de los noventa, encontré que la distribución de resultados para alumnos de

undécimo grado sin discapacidades se asemejaba mucho a una curva de campana, pero la distribución de los estudiantes con discapacidades mostraba un fuerte sesgo: los resultados se concentraban en la parte inferior del rango y una cola delgada se extendía hacia el rango de mayor puntaje. La prueba era demasiado difícil para algunos de los estudiantes con discapacidades.*

Para resolver nuestro problema podemos sacar ventaja del hecho de que los resultados de nuestro ejemplo se ajustan a la curva de campana, esto es, señalar los resultados en una escala que no sea arbitraria y que nos permita comparar los resultados de una prueba con los de otra. Lo haremos con nuestros datos reales de lectura.

Primero, haremos uso de la desviación estándar, que mide cómo se dispersan los resultados. Por el momento no es importante la definición técnica de la desviación estándar;† lo que importa es, que dado que los resultados (o cualquier otra medición) se ajustan a la curva de campana, sabemos qué proporción de la distribución cae dentro de cualquier rango definido en términos de desviaciones estándar. Por ejemplo, si se establece un corte en una desviación estándar por arriba de la media, 84 por ciento de todos los resultados caerán por debajo de esa línea y 16 por ciento por arriba de ella. Esto es simétrico: si se establece el corte de una desviación estándar por debajo de la media, 16 por ciento de los resultados caerán por debajo de ella. Si establece el corte en dos desviaciones estándar por arriba de la media, alrededor de 95 por ciento de los resultados caerán por debajo del corte. Y esto será así sin importar la escala original, en la medida que la distribución se ajuste aproximadamente a la curva de campana. Por lo tanto, en la medida que la distribución se ajuste aproximada-

* Las calificaciones de equivalencia de grado en los primeros grados son otra excepción; debido a su construcción se sesgan a la derecha.

† La desviación estándar se basa en las diferencias de cada observación respecto a la media. En concreto, es la raíz cuadrada del cuadrado promedio de la desviación respecto de la media.

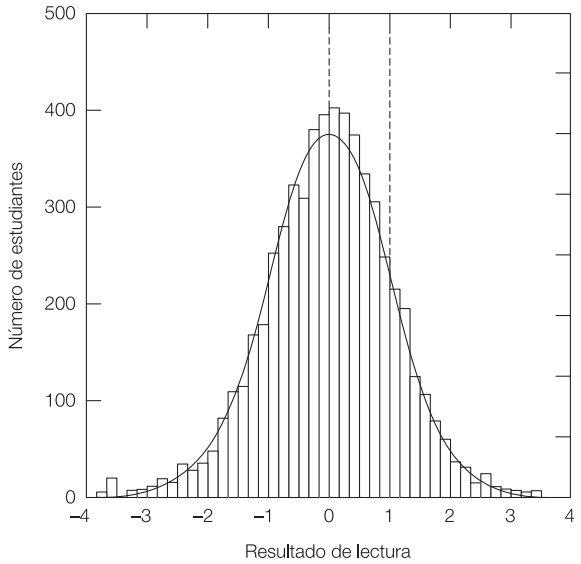
mente a la curva de campana es posible emplear desviaciones estándar para colocar los resultados obtenidos en una prueba en una escala con un significado claro y bien entendido. Si yo digo que un individuo obtuvo una puntuación de 633 en la porción matemática de la prueba SAT, sólo la gente muy familiarizada con esa prueba sabrá lo que significa. Por otro lado, si digo que la puntuación del estudiante se sitúa a una desviación estándar por arriba de la media en la prueba SAT, cualquiera que esté familiarizado con las desviaciones estándar sabe que eso significa que el estudiante superó a alrededor de 84 por ciento de los alumnos que presentaron la prueba.

Nuestros datos estatales de lectura en la escala original mostrados en las figuras 5.1 y 5.2 tienen una media de alrededor de 500 y una desviación estándar más o menos de 100. Esta es la razón por la que elegí las puntuaciones de 500 y 600 para las líneas punteadas de la ilustración: son la media y una desviación estándar por arriba de la media. Si quisiera contar a los estudiantes representados por las barras arriba de 600, encontraría que aproximadamente 16 por ciento calificó en ese rango. El histograma es un poco desigual, por lo que este porcentaje no es exacto, pero la distribución normal nos ofrece una aproximación bastante buena.

Los psicómetras y otros científicos sociales suelen sacar ventaja de esas propiedades de la curva de campana convirtiendo los resultados obtenidos de la prueba y otras variables a una escala *estandarizada*, que es una escala con una media de cero y una desviación estándar de uno.* Al estandarizar los resultados de lectura obtenemos la figura 5.3. La distribución muestra exactamente la misma forma que en las figuras anteriores, pero se modificó la escala en el fondo, de modo que ahora la media es 0, la desviación

* Esta escala suele denominarse una escala *estándar*, pero aquí utilizaremos el término *estandarizada* para distinguirla de otras escalas relacionadas, revisadas en el capítulo 8, que también se conocen como *escalas estándar*.

Figura 5.3. La distribución de los resultados de lectura de un estado no identificado sobre una escala estandarizada



estándar es 1 y las líneas punteadas verticales están en 0 y en 1 en lugar de 500 y 600. El resultado de cada estudiante ahora se representa como el número de desviaciones estándar (o la fracción de una desviación estándar), que se encuentra del nuevo resultado promedio total de cero.

Una vez que hemos transformado los datos de los resultados a esta escala estandarizada, hay varias formas de darles sentido. Podemos comparar las tendencias de una prueba con las de otra; comparar un cambio o diferencia en los resultados en una determinada prueba con otras diferencias, como las diferencias promedio entre los resultados de los estudiantes negros y blancos en Estados Unidos o entre ese país y otras naciones con altos resultados, como Corea y Singapur (ambos casos se analizan más adelante). También podemos aprovechar nuestra comprensión de la curva de campana para expresar diferencias en términos más completos

y sencillos de entender. Por ejemplo, digamos que la línea vertical superior de la figura 5.3 representa la calificación promedio de otro grupo de estudiantes. Eso nos mostraría que el promedio de ese otro grupo supera al promedio de nuestro grupo en una desviación estándar completa. (Para anticipar un poco de la trama, ahí es más o menos donde caen los resultados promedio de Japón y Corea en la distribución de los resultados obtenidos por Estados Unidos en algunas pruebas de matemáticas.) Al inicio, sólo sabríamos que este grupo obtuvo un promedio de 600 en nuestra escala original, lo que nos diría muy poco. Pero ahora que estamos al tanto de que su promedio está una desviación estándar por arriba de nuestra media, sabemos que sólo alrededor de 16 por ciento de nuestros alumnos califican por arriba de su promedio, lo que constituye claramente una enorme diferencia en el desempeño.

Este ejemplo muestra otra manera de usar los resultados estandarizados, que consiste en replantearlos en términos de *rangos percentiles*. El rango percentil es el porcentaje de resultados que caen por debajo de cualquier resultado dado. Suponga que aplicamos la prueba de vocabulario de 40 reactivos de la que hablamos en el capítulo 2 a un grupo grande de personas y que 75 por ciento de ellas respondieron correctamente a menos de 30 reactivos. Se diría entonces que una estudiante que respondió de manera correcta 30 reactivos tiene un rango percentil de 75, lo que quiere decir que superó a 75 por ciento del grupo de referencia. Como muestran los ejemplos presentados antes, la relación entre los resultados estandarizados y los rangos percentiles es fija y se conoce bien, por lo que una vez que se han estandarizado los resultados es muy sencillo convertirlos a rangos percentiles.*

* Una tabla que se encuentra prácticamente en todos los libros de estadística y que es fácil obtener en internet, denominada "tabla normal estándar" o "tabla *z*" muestra la representación de los resultados estandarizados como rangos percentiles.

Si las escalas estandarizadas son tan prácticas, ¿por qué no se usan para presentar al público los resultados obtenidos en las pruebas? Porque la mayoría de las personas no especializadas no toleran los resultados fraccionales y, menos aún, los resultados o calificaciones negativos. Imagine que un padre recibe un reporte escolar de su hija que dice: "Su hija obtuvo una calificación de 0.50, que la coloca muy por arriba del promedio". Un resultado estandarizado de 0.50 *está* muy por arriba del promedio (si los resultados se ajustan exactamente a la curva de campana, representa un rango percentil de 69), pero ciertamente no lo parece. Es muy fácil imaginarse la respuesta del padre preocupado porque no entiende cómo es que su hija obtuvo un resultado de un medio y aún así terminó por arriba del promedio. ¿La mitad del grupo no tuvo aciertos? O peor todavía, imagine un reporte que diga: "La calificación de su hijo en nuestra prueba de matemáticas fue de -.25" ¿Menos de cero? No me gustaría ser ese niño en la mesa del comedor la noche en que el reporte llegara a su casa. De modo que en muchos casos, los psicómetras empiezan con una escala estandarizada (muchos métodos actuales de escalamiento la producen automáticamente) y luego la convierten a alguna otra escala que consideran que será más agradable, con una desviación estándar más grande para evitar las calificaciones o resultados fraccionales y una media lo bastante elevada para que nadie reciba una calificación o resultado negativo.

Tendencias en el logro de los estudiantes estadounidenses

Con las escalas estandarizadas a la mano, podemos comparar los resultados de muchas pruebas diferentes para obtener una perspectiva general de las tendencias en el desempeño de los estudiantes estadounidenses.

Durante la década de los sesenta, los resultados obtenidos por los educandos estadounidenses empezaron a caer. Los resultados en la prueba SAT empezaron a disminuir en el año escolar de 1963; unos años más tarde sucedió lo mismo en la prueba ACT y en diversos momentos de la década también se observó un deterioro en los resultados obtenidos en otras pruebas. Aunque la atención del público estaba concentrada en unas cuantas pruebas, en particular la prueba SAT, el descenso fue sorprendentemente generalizado. Con algunas excepciones, se presentó en las pruebas de admisión a la universidad, la Evaluación Nacional del Progreso Educativo, otros estudios representativos a escala nacional que se realizaron bajo contrato con el gobierno federal, pruebas estatales, diversos estudios especiales y en los datos de normas nacionales obtenidos de pruebas comerciales de logro (datos de las evaluaciones nacionales periódicas realizadas con la intención de establecer normas nacionales para reportar calificaciones). También ocurrió en Canadá y en las escuelas públicas y privadas de Estados Unidos.

Es difícil exagerar el impacto de esta disminución en los resultados. Aunque sólo se reconoció de manera general en la década de los ochenta, después de que había terminado, tuvo efectos profundos en la opinión que se tenía en Estados Unidos de su sistema educativo. Por ejemplo, fue el principal foco de atención de *Una nación en riesgo*,[1] un informe de enorme influencia que se presentó en 1983, el cual empezaba con una exposición alarmante:

> Nuestra nación está en riesgo. Nuestra supremacía, alguna vez indiscutible en el comercio, la industria, la ciencia y la innovación tecnológica, nos está siendo arrebatada por los competidores de todo el mundo. Este reporte sólo se interesa en una de las muchas causas y dimensiones del problema, que es la que sustenta la prosperidad, la seguridad y la urbanidad estadounidense. Informamos al pueblo de Estados Unidos que si bien es justificado nuestro

> orgullo por los logros y contribuciones históricas de nuestras escuelas y universidades a los Estados Unidos y al bienestar de su gente, en la actualidad los cimientos educativos de nuestra sociedad están siendo erosionados por una corriente creciente de mediocridad que amenaza nuestro futuro como nación y como pueblo.

El informe presentaba 13 diferentes "indicadores de riesgo" para apoyar esta sombría valoración; excepto tres, todos se basaban en los resultados obtenidos en las pruebas. La lista era encabezada por los pobres resultados obtenidos por los estudiantes estadounidenses en las pruebas internacionales, pero más o menos la mitad de los indicadores eran disminuciones en los resultados.

La preocupación por la calidad de la educación estadounidense, a la que contribuyó de manera importante la disminución de los resultados, no se mitigó en las décadas transcurridas. En sí mismo, el declive ha perdido gran parte de su prominencia y ha cedido el paso a una mayor atención a las comparaciones internacionales que ahora son más frecuentes y a los datos más recientes sobre las tendencias en los resultados dentro de Estados Unidos. Sin embargo, la disminución de los resultados no ha desaparecido del todo de la óptica y todavía reaparece aquí y allá como base para juzgar el desempeño actual. Por ejemplo, un artículo reciente declaraba:

> Como consecuencia de *Una nación en riesgo*, los educadores se comprometieron a enfocarse de nuevo en el logro del estudiante. Dos décadas más tarde, es poco el progreso realizado... Muchas áreas de la vida estadounidense han mejorado durante las dos décadas pasadas, salvo, al parecer, el sistema de educación primaria y media superior. Datos de diversas fuentes –la prueba SAT, la Evaluación Nacional del Progreso Educativo (*National Assessment of Educational Progress*, NAEP) y comparaciones internacionales como el Tercer Estudio Internacional de Matemáticas y Ciencia

> (*Third International Mathematics and Science Study*, TIMSS)– revelan la misma tendencia: a pesar de 20 años de agitación y reforma, provocada en gran medida por el reporte de *Riesgo*, en el mejor de los casos el logro de los estudiantes se ha estancado si no es que declinado.[2]

Es posible encontrar descripciones menos negativas y ciertamente menos hiperbólicas de los datos, pero en general, desde la publicación de *Una nación en riesgo*, en el debate público ha predominado una visión pesimista de los resultados de las pruebas.

Aunque no es precisamente una base para la complacencia, la evidencia es menos negativa de lo que sugiere la cita. Los datos de las pruebas ofrecen buenas y malas noticias y el cuadro que pintan es complejo, con algunos ángulos interesantes. Además, hay algunas complicaciones que es necesario tener en mente para poder interpretar las tendencias con precisión.

¿Qué tan grande fue esa disminución en los resultados que puso a nuestra nación en riesgo? No hay una sola respuesta, pero a grandes rasgos y considerando muchas fuentes distintas de datos, sería justo decir que la caída fue "moderadamente grande". El tamaño del declive variaba considerablemente de una prueba a otra, en ocasiones por razones comprensibles pero en muchos casos de manera inexplicable. La mayor parte de las pruebas mostraron una caída total de entre 0.25 y 0.40 desviaciones estándar en el curso de la disminución. El uso de esta escala nos permite aclarar el tamaño del deterioro al estimar cómo se compararían los estudiantes examinados al finalizar el declive con sus pares antes de que empezara la caída. Por ejemplo, consideremos al estudiante que alcanzó la mediana –es decir, el estudiante que superó a la mitad del grupo examinado y que por ende tenía un rango percentil de 50– al final de una disminución de 0.35 desviaciones estándar, que era una caída bastante típica. El resultado que colocó a este estudiante en la mediana

al final del declive hubiera sido suficiente para alcanzar sólo el rango percentil 36 antes de que los resultados empezaran a caer.

Uno de los mayores deterioros se observó en la parte verbal de la prueba SAT, donde los resultados cayeron casi la mitad de una desviación estándar. Esta tendencia particular tuvo una influencia desmedida en el debate público debido a la prominencia de dicha prueba, pero era engañosa y la razón de que lo fuera (los cambios en la composición del grupo de estudiantes que presentaron la prueba) tuvo implicaciones mucho mayores.

Para ilustrar el problema general que surge cuando cambia la composición del grupo examinado, considere a los estudiantes con necesidades especiales. Desde hace más de una década, las políticas federales y estatales han presionado a los distritos escolares para incluir a más estudiantes con discapacidades y a los que no dominan el inglés en las evaluaciones regulares que se aplican a otros estudiantes. El desempeño de los alumnos en esos grupos es muy variable, pero en promedio tienden a obtener resultados más bajos que la población estudiantil general. Considere ahora un distrito hipotético en que la eficacia de la educación permaneciera constante y en que las características de la población estudiantil total no fueran cambiadas en un periodo de 10 años. Suponga que este distrito al principio excluyó de la evaluación a muchos estudiantes con discapacidades y con un dominio limitado del inglés, pero que en el curso de la década se incrementó con rapidez la proporción incluida. Los resultados promedio habrían tenido una caída modesta. ¿Qué significaría esta caída en los resultados? No significaría que las escuelas se hubiesen vuelto menos eficaces o que el aprendizaje de los estudiantes, por cualquier razón, fuese menor. La disminución no señalaría otra cosa que un cambio en la selección de los estudiantes examinados.

A esas variaciones se les llama *efectos de la composición* –cambios en el desempeño que surgen de las modificaciones en la

composición del grupo evaluado. En general, si los resultados promedio de los subgrupos que están creciendo son considerablemente diferentes a los resultados de los que constituyen la parte decreciente del grupo, el resultado será un cambio en la puntuación promedio general que surge simplemente de esas tendencias en la composición. Como ilustra el ejemplo, se requiere cierta cautela para dar sentido a las tendencias frente a los efectos de la composición: algunas interpretaciones de las tendencias estarán justificadas, pero es probable que otras no.*

Las características de la población estudiantil estadounidense han cambiado de formas tales que han tenido impacto en los resultados obtenidos en las pruebas. Por ejemplo, la inmigración ha incrementado la proporción de alumnos con un dominio limitado del inglés. Las tendencias en las tasas de graduación y de deserción han modificado la composición de la población estudiantil en la secundaria a medida que un número creciente de alumnos que antes habrían desertado ahora permanecen en la escuela y, por ende, han sido examinados. Por ejemplo, entre 1970 y 1985 se observó una disminución sorprendente y constante en la tasa de deserción de preparatoria de los estudiantes negros (medida por uno de los numerosos indicadores de deserción, la llamada

* Los efectos de la composición son un caso especial de la *paradoja de Simpson*, el hecho de que los efectos del tratamiento u otros cambios evidentes dentro de los grupos pueden ser contradictorios e incluso oponerse a los efectos que se presentan cuando se combinan los grupos. Por ejemplo, un estudio puede mostrar que un tratamiento médico es eficaz para hombres y mujeres, pero cuando se combinan los resultados de los géneros parece mostrar de manera engañosa que es ineficaz. La paradoja de Simpson se ha discutido durante varios años en la estadística (recibió su nombre por un artículo que publicó E. H. Simpson en el *Journal of the Royal Statistical Society* en 1951 y antes de eso había sido reconocida en las publicaciones estadísticas tal vez durante medio siglo) y ha sido encontrada en incontables estudios médicos y de las ciencias sociales. No obstante, el fenómeno no se ha convertido en algo trivial y los lectores no especializados en los análisis estadísticos a menudo se confunden.

tasa del estatus de deserción, que es la proporción de individuos, entre 18 y 24 años, que no se han graduado de la preparatoria ni están actualmente inscritos en la escuela).

El efecto de los cambios en la composición puede agravarse cuando la presentación de la prueba es voluntaria, y la disminución en los resultados de la prueba SAT fue empeorada por un cambio importante en la composición: un incremento considerable en la proporción de sustentantes de la prueba SAT provenientes de grupos que históricamente han recibido bajas calificaciones. A medida que la asistencia a la universidad se hizo más común, aumentó la proporción de graduados de la preparatoria que decidieron presentar pruebas de admisión y muchos de los recién agregados a las listas eran estudiantes de bajas calificaciones. Este asunto fue estudiado con considerable detalle por el Consejo de Exámenes de Ingreso a la Universidad en la década de los setenta, y la investigación demostró con claridad que una parte considerable de la caída en los resultados de la prueba SAT era consecuencia de este cambio en la composición. Si las características del grupo que presentó la prueba se hubiesen mantenido constantes, la disminución habría sido mucho menor.

Aunque los efectos de la composición pueden ser un problema de particular importancia cuando los estudiantes deciden si van a presentar una prueba, también pueden afectar las tendencias de otros instrumentos cuando la composición de toda la cohorte cambia a lo largo del tiempo. En las décadas recientes, la composición de la población estudiantil de Estados Unidos ha cambiado de muchas formas y esos cambios han tenido un modesto efecto negativo sobre los resultados, ya sea agravando el deterioro o disminuyendo el incremento que se habría observado de permanecer estable la población. Un ejemplo que se ha divulgado mucho en la actualidad es el rápido crecimiento en la proporción de estudiantes que no tienen un dominio pleno del inglés. Con base en una encuesta anual de hogares, la Encuesta de la Población Actual

(Current Population Survey, CPS) el Buró del Censo calcula que entre 1979 y 2004 el porcentaje de niños entre cinco y 17 años que hablaban en casa un idioma distinto al inglés se incrementó de 9 a 19 por ciento. En el mismo periodo, el porcentaje de niños de ese mismo grupo de edad que hablaban el inglés con dificultad casi se duplicó, de 2.8 a 5.3 por ciento.[3] Desde 1978 los estudiantes afroamericanos han representado aproximadamente 16 por ciento de los alumnos de escuelas públicas, pero la inscripción de los estudiantes hispanos, que de manera histórica también han recibido resultados promedio considerablemente menores a los blancos no hispanos, aumentó de cerca de 7 a 19 por ciento.

Se requiere de cierto cuidado para interpretar las tendencias de los resultados cuando está presente un cambio en la composición, pero en los pasados 40 años, la mayoría de los comentaristas no tuvieron conocimiento de esa situación o tal vez no les inquietaba. Veamos de nuevo el caso de la prueba SAT. La disminución de los resultados en la parte verbal de esa prueba que terminó en 1980 proporcionó una descripción razonable del desempeño del grupo cambiante de estudiantes que presentaban la prueba, la cual quizá fue relevante para los funcionarios encargados del departamento de admisión a la universidad, quienes tenían razones para preocuparse por las capacidades de cualquier subconjunto de graduados que solicitaran admisión. Sin embargo, esta tendencia de los resultados, de no ser ajustada a los cambios en la composición, pintaba un cuadro sesgado del cambio en el desempeño de todos los graduados. Y eso fue lo que hizo la mayoría de los comentaristas de los resultados en la prueba SAT cuando las citaban como una descripción de la disminución del desempeño de los graduados de preparatoria en Estados Unidos o cuando –llevando la cuestión más lejos de lo justificado– las usaban para hacer inferencias sobre la calidad de las preparatorias estadounidenses. Para tales fines, la simple tendencia en los resultados obtenidos en la prueba SAT exageraba

la disminución del desempeño entre todos los graduados de preparatoria o, para ser más precisos, exageraba la disminución de los resultados en la prueba SAT que se habría encontrado si todos los graduados hubiesen presentado el examen en lugar de los estudiantes cambiantes, autoseleccionados, que en realidad lo hicieron.

La terminación del declive –tanto en su momento como en su explicación– fue tema de controversia entre los conocedores de las tendencias. La mayoría de los comentaristas se enfocaron sobre todo en la prueba SAT, que tocó fondo en 1980 y empezó a mostrar indicios de un pequeño incremento dos años más tarde. Esto los llevó a la búsqueda de las posibles causas, educativas o sociales, que ocurrían por esa época. Algunos comentaristas sugirieron incluso que los cambios conservadores, culturales y políticos, que acaecían en el país en el momento en que los resultados en la prueba SAT tocaron fondo, reflejados en la elección de Ronald Reagan en 1980, detuvieron la disminución de los resultados. Estaban buscando en el lugar equivocado. El logro de los estudiantes que se gradúan de la preparatoria representa los efectos acumulados de 12 años de estudios, por lo que algunos de los factores que contribuyeron a la finalización del deterioro en los resultados pueden haber precedido por mucho tiempo al hundimiento de la SAT. Como dijo en broma H. D. Hoover, de los programas de evaluación de Iowa: "Si la gente va a ser lo bastante tonta para atribuir algo como esto a un presidente, tendría que darle crédito a Jerry Ford".

Y si se examinan más de cerca, los datos muestran justamente eso: el deterioro en los resultados terminó antes en los grados inferiores, durante la administración de Ford. Se observó una variación considerable de un grupo a otro y de una prueba a otra, pero si se ve una amplia variedad de datos, el punto más bajo de los resultados por lo general ocurrió entre 1974 y 1980. Los estudiantes cuyos resultados representaban el punto bajo parecen haber sido los que nacieron más o menos en 1962 y 1963. El final de esta dis-

minución apareció por primera vez a mediados de la década de los setenta (cuando esas cohortes de nacimiento llegaron a la escuela secundaria) y progresó a los grados superiores en los años posteriores a medida que esos estudiantes se abrían camino en la escuela. Este es un ejemplo de lo que los científicos sociales llaman un *efecto de la cohorte*: un cambio vinculado a una determinada cohorte de nacimiento y que la sigue a medida que crece. Los efectos de la cohorte son comunes (por ejemplo, ocurren en medicina cuando ciertas cohortes de nacimiento son expuestas a determinadas condiciones o intervenciones médicas) pero no ocuparon un lugar destacado en los debates sobre la política educativa de la época. Más bien, la mayoría de los comentaristas veían las tendencias en términos de los *efectos del periodo* (cambios vinculados a un momento particular, como las políticas impuestas a las escuelas en un periodo específico) que afectan a muchas cohortes al mismo tiempo.

A principios de la década de los ochenta, después de leer un trabajo en que yo describía esos patrones, recibí la visita del redactor principal de un periódico importante quien me preguntó: "¿No indica esto que las críticas educativas que atribuyeron la finalización del declive a las políticas conservadoras de los ochenta están llenas de basura?". Bueno, sí, le dije, pero las implicaciones generales son más importantes. Primero, es esencial tener en mente que la educación es un proceso de largo plazo, cuyos efectos se acumulan a lo largo del tiempo. Es riesgoso tratar de explicar los cambios en el desempeño haciendo referencia solamente a los cambios recientes en las políticas o las prácticas educativas. Segundo (como explicaré en el siguiente capítulo), esos patrones de cohorte sugieren que tanto los factores sociales como los educativos desempeñaron un papel importante en la generación del deterioro en los resultados y el repunte posterior.

¿Qué ha sucedido desde los años ochenta? Podría esperarse que con el enorme incremento en la cantidad de evaluaciones

realizadas en los años recientes, conoceríamos más sobre las tendencias actuales que sobre el deterioro de las décadas de los sesenta y setenta. Irónicamente ha sucedido lo contrario. Aunque ahora contamos con muchos más datos que hace 20 o 30 años, disponemos de menos fuentes de datos confiables. La razón es sencilla: el incremento en la evaluación ha sido acompañado por un impresionante recrudecimiento de las consecuencias ligadas a los resultados. A su vez, eso ha creado incentivos para tomar atajos (diversas formas de preparación inadecuada para la prueba, incluyendo las trampas descaradas) que pueden producir una inflación importante de los resultados, lo que hace que tales tendencias se vuelvan engañosas o incluso carentes de sentido. Este problema, que se analiza con detalle en el capítulo 10, ha ocasionado ganancias exageradas en los resultados en algunas pruebas de alto impacto que son mucho más marcadas que los incrementos en pruebas de poco impacto como la NAEP. En consecuencia, para discernir las tendencias recientes tenemos que confiar sobre todo en esa prueba.

Si bien en muchos sentidos la Evaluación Nacional del Progreso Educativo es ideal para ese propósito (está cuidadosamente diseñada para medir tendencias, refleja cierto grado de consenso sobre lo que los estudiantes deben conocer y su cobertura de los dominios medidos es inusualmente amplia), es muy arriesgado depender tanto de una sola fuente, no importa lo buena que sea. Incluso las pruebas bien diseñadas a menudo proporcionan visiones muy distintas de las tendencias debido a las diferencias en su contenido y otros aspectos de su diseño. En efecto, como se observó en el capítulo 1, las dos Evaluaciones Nacionales del Progreso Educativo (la "principal" que se usa para informar resultados detallados a nivel nacional y estatal y el estudio más pequeño que se emplea para evaluar las tendencias a largo plazo) en ocasiones han mostrado tendencias considerablemente diferentes. La NAEP puede proporcionar un amplio panorama de las tendencias,

pero debemos tener cuidado de no confiar demasiado en los hallazgos detallados, como el tamaño preciso de los cambios a lo largo del tiempo o de las diferencias entre grupos.

La NAEP presenta una imagen desalentadora de las tendencias en lectura. En secundaria y preparatoria no se ha observado un cambio importante en los resultados de lectura durante más de 30 años. Las tendencias en la escuela primaria son un poco mejores, con aumentos durante la década de los setenta y aumentos muy pequeños desde 1999 o 2000. Pero el desempeño actual en la escuela primaria es apenas un poco mayor que en 1980.

La historia es muy diferente en el caso de las matemáticas: se han observado aumentos grandes y rápidos en los resultados de matemáticas de la escuela primaria y mejoras considerables en la escuela secundaria. Los resultados de los niños de nueve años en la evaluación de la tendencia de largo plazo empezaron a elevarse después de 1982 y ahora se encuentran en el nivel más alto que se ha registrado nunca. El aumento fue tan rápido que, a pesar de un periodo de estancamiento en la década de los noventa, el resultado promedio en 2007 fue 0.84 desviaciones estándar más alta que en 1982. Esto es mostrado por la línea continua de la figura 5.4, que establece el primer año de cada tendencia en una media de cero y luego muestra los incrementos como una fracción de una desviación estándar. La NAEP principal puede utilizarse para medir tendencias sólo por un periodo más corto, pero muestra una mejora incluso más rápida. El resultado promedio para los estudiantes de cuarto grado en esta evaluación se incrementó en alrededor de dos tercios de una desviación estándar, pero en un periodo de sólo 13 años, de 1990 a 2003 (véase la línea punteada de la figura 5.4).

Una forma de poner en perspectiva este incremento consiste en compararlo con el deterioro que desató la preocupación por las condiciones de la educación estadounidense. Los incrementos obtenidos por las escuelas primarias en los resultados de matemáticas en

Figura 5.4. Aumento acumulativo en los resultados de matemáticas en la escuela primaria, NAEP

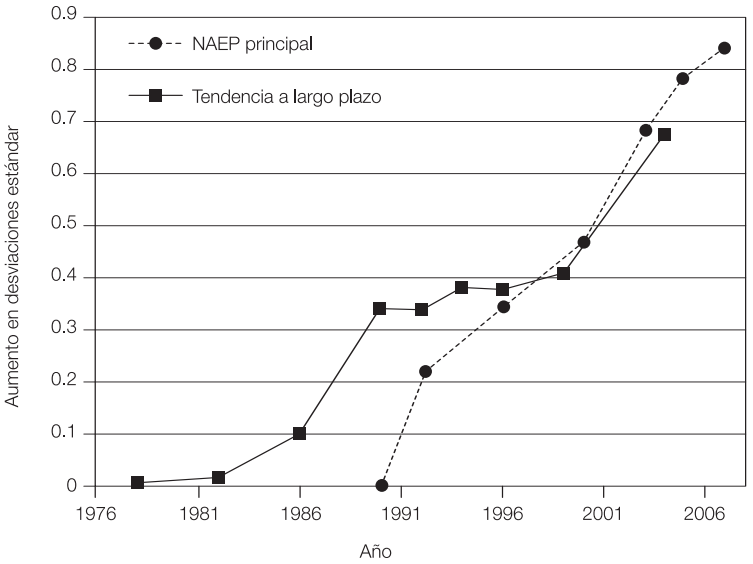
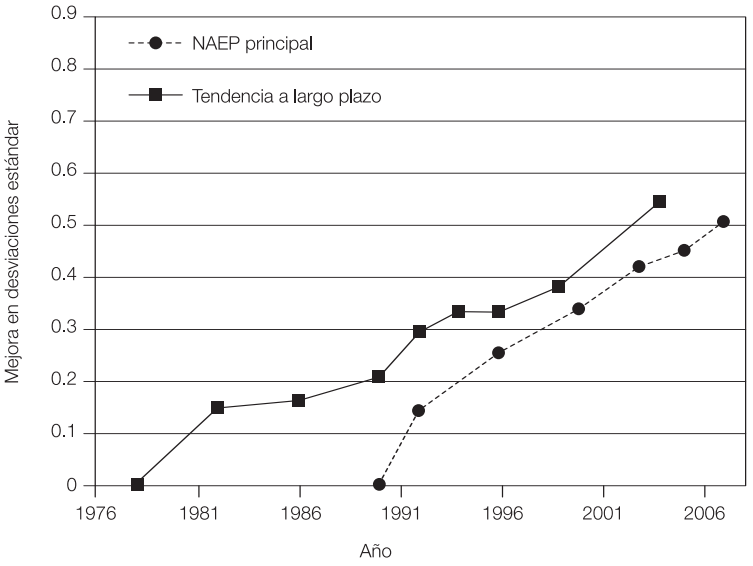


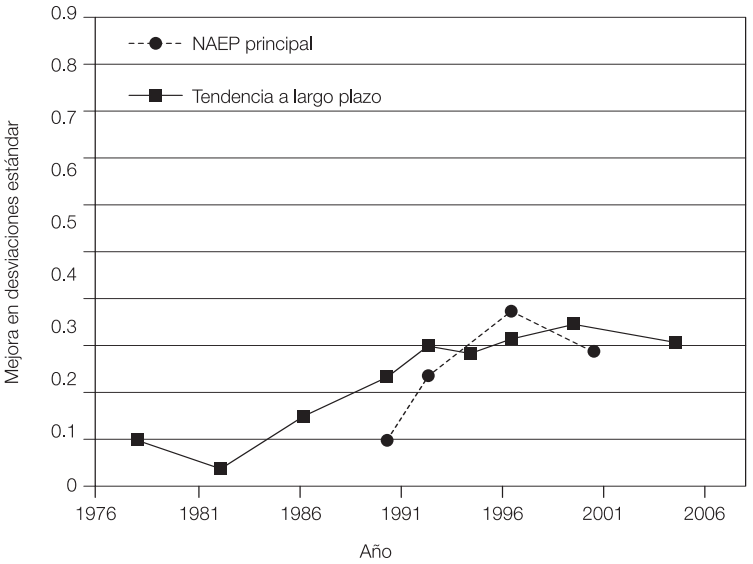
Figura 5.5. Aumento acumulativo en las calificaciones de matemáticas en la escuela secundaria (*middle-school*), NAEP



la NAEP son mucho mayores que la mayor parte de las disminuciones registradas en los resultados y son casi el doble de la caída típica. Además, la tasa anual de cambio en los resultados obtenidos en la NAEP de largo plazo es más o menos comparable a la observada durante el declive, mientras que el incremento en la NAEP principal es mucho más rápido. Otros dos estándares útiles de comparación son las brechas en los resultados promedio entre los estudiantes afroamericanos y blancos y entre Estados Unidos y Japón en las evaluaciones internacionales. La mejora de 0.84 desviaciones estándar en los resultados de matemáticas de la NAEP en las escuelas primarias cae en el rango encontrado para la diferencia promedio entre afroamericanos y blancos, y es similar al tamaño de la brecha entre Estados Unidos y Japón en matemáticas de octavo grado. Incluso los críticos que insisten en que los resultados no han demostrado mejoras apreciables considerarían que las brechas entre afroamericanos y blancos y entre Estados Unidos y Japón son grandes, como en realidad lo son. De igual modo, también la mejora en el desempeño de los alumnos de primaria en matemáticas ha sido considerable.

La mejora en los resultados de matemáticas en la secundaria fue menos notable, pero también fue aceptable y colocó al desempeño actual en el nivel más alto desde el inicio de las evaluaciones. Los resultados promedio de los jóvenes de 13 años en la evaluación de las tendencias a largo plazo exhibieron un incremento más lento pero menos errático que las mejoras mostradas por los niños de nueve años: un incremento de apenas la mitad de una desviación estándar en un periodo de 26 años, de 1976 a 2004 (figura 5.5). El incremento entre los alumnos de octavo grado en la evaluación principal fue de alrededor de media desviación estándar en un periodo más corto, los 17 años de 1990 a 2007. Una vez más, esas mejoras son mayores que la disminución típica en los resultados que les precedió.

Figura 5.6. Aumento acumulativo en los resultados de matemáticas en la preparatoria, NAEP



La mala noticia es que las tendencias en el nivel de la preparatoria son mucho menos alentadoras. En las dos evaluaciones NAEP se han entremezclado incrementos más lentos con breves recaídas entre los alumnos de 17 años de doceavo grado y en ambas evaluaciones el incremento total ha sido apenas de 0.2 desviaciones estándar (figura 5.6). No obstante, incluso esta mejora modesta se encuentra en el rango de cambios observados durante el periodo de disminución de los resultados.*

La prueba SAT pinta una visión más optimista de las tendencias al final de la preparatoria que la NAEP. La prueba SAT es menos útil que la NAEP para valorar el logro total de los estudiantes

* Cuando se estaba escribiendo esto, los datos más recientes de la NAEP principal en el desempeño en matemáticas para los alumnos de doceavo grado eran del 2000.

estadounidenses porque no se diseñó para reflejar currículos comunes de preparatoria y es presentada por un grupo autoseleccionado y cambiante de alumnos; no obstante, amerita que se tome en cuenta y ciertamente juega un papel destacado en el debate público. Los resultados en la porción verbal de la prueba SAT se estancaron desde que tocaron fondo, con pequeñas fluctuaciones, pero terminaron con apenas seis puntos (0.05 desviaciones estándar) más altas en 2005 que en 1980.

En contraste, los resultados promedio en la parte de matemáticas de la prueba SAT han mostrado un incremento constante, pero lento, desde el final del declive. En 2005 el resultado promedio fue 28 puntos (0.24 desviaciones estándar) mayor que en 1980 y alcanzó su punto más alto desde 1967.

Los cambios en la composición han atenuado un poco esas mejoras. La disminución de las mejoras en la prueba SAT se manifiesta en el hecho de que las tendencias *en el interior* de los grupos raciales y étnicos a menudo eran más favorables que la tendencia total. Los cambios en la composición racial y étnica del grupo que presentaba la prueba redujo en parte el incremento que se habría presentado si la composición del grupo se hubiera mantenido fija. Los efectos de la composición han afectado también las tendencias de la NAEP, aunque de manera más modesta porque en cada ocasión el examen es presentado por una muestra representativa. Por ejemplo, la NAEP reporta que la mejora en el desempeño en matemáticas de los alumnos de 17 años entre 1973 y 2004 (un periodo que abarca parte de la disminución y del incremento posterior) era demasiado pequeña para ser estadísticamente significativa. Es decir, el incremento era tan pequeño que podría ser resultado del azar más que de mejoras reales en el aprendizaje. (En el capítulo 7 se explican la significancia estadística y otros conceptos relacionados.) No obstante, las mejoras fueron estadísticamente significativas para blancos, negros e hispanos tomados por

separado –aunque las tendencias en los grupos más pequeños por lo general son estadísticamente menos significativas.[4] La explicación para esta aparente paradoja es que los incrementos dentro de cada grupo fueron reducidos en parte por un cambio en la mezcla de estudiantes examinados, en que los grupos raciales y étnicos de menores resultados aumentaron su participación en la población escolar total.

Es claro que esos datos no justifican la conclusión de que "a pesar de 20 años de agitación y reforma, provocadas en gran medida por el reporte de *Riesgo*, en el mejor de los casos el logro de los estudiantes se ha estancado si no es que declinado". Aun así, existen buenas razones para preocuparse, entre las que destaca el deterioro de las mejoras recientes a medida que los estudiantes avanzan en la escuela. Esos datos muestran las mismas cohortes de estudiantes en diferentes años (no los mismos estudiantes en cada ocasión sino muestras nacionalmente representativas). La cohorte evaluada con la NAEP a los nueve años en 1982 reaparece en los datos de 1986 como alumnos de 13 años y en los datos de 1990 como participantes de 17 años. Por lo tanto, si los estudiantes estuvieran manteniendo sus incrementos iniciales, debería apreciarse que las mejoras de la escuela primaria son similares a las obtenidas cuatro años más tarde en la secundaria y a las que se ven en la preparatoria cuatro años después. Esto es lo que se vio al final del declive, pero no es lo que vemos con el aumento posterior de los resultados. Dado que en última instancia estamos interesados en el conocimiento, las habilidades y disposiciones que los estudiantes acumulan en el curso de su educación y llevan consigo a la adultez, se justifica entonces la profunda preocupación de los educadores y los críticos del sistema por este deterioro.

¿Qué tan variable es el desempeño de los estudiantes?

La mayoría de los reportes de los resultados agregados (los de escuelas, estados, distritos e incluso países) se enfocan en la competencia tipo carrera de caballos: ¿quién obtiene un resultado mayor que quién? A menudo, los diarios informan sólo un resultado promedio o el porcentaje de alumnos que superan algún estándar de desempeño, junto con el resultado de escuelas, estados o incluso países.

La información sobre la *variabilidad* es al menos tan importante como esas simples diferencias en los niveles de desempeño. ¿Cuál es la dispersión típica del desempeño de los estudiantes en las pruebas de logro? ¿Qué tan constantes y tan grandes son las diferencias entre grupos socialmente importantes, por ejemplo, entre grupos raciales y étnicos, entre hombres y mujeres, y entre alumnos de familias con bajos y con altos ingresos?

Es posible que en Estados Unidos la brecha entre los estudiantes afroamericanos y blancos sea el estadístico más analizado en relación con la variabilidad. De manera tradicional, esto se ha presentado como la *diferencia promedio* (la diferencia en los resultados promedio de cada grupo), pero en los años recientes se ha vuelto común que se reporte como una diferencia en el porcentaje de alumnos que alcanzan cierto estándar de desempeño. En este caso me concentro en las diferencias promedio porque es muy difícil interpretar los cambios en el porcentaje de alumnos que alcanzan un estándar; estos últimos combinan la diferencia de desempeño entre grupos con el nivel en que se estableció el estándar.

La diferencia promedio entre blancos y afroamericanos ha variado considerablemente de una fuente de datos a otra debido a las tendencias en el tamaño de la brecha a lo largo del tiempo, las diferencias entre las pruebas empleadas y las diferencias entre las muestras seleccionadas para la evaluación. No obstante, la diferencia ha sido grande en todos los estudios confiables de grupos

representativos de estudiantes en edad escolar. La mayor parte de las estimaciones de la diferencia se ubican en el rango de 0.8 a 1.1 desviaciones estándar, aunque algunas han sido un poco más pequeñas. Para entender lo que esto significa veamos un ejemplo: la diferencia promedio entre los afroamericanos y los blancos en la evaluación principal de la NAEP realizada en 2000 para las matemáticas de octavo grado, que fue de 1.06 desviaciones estándar. Esta diferencia estandarizada indica que el estudiante negro localizado en la mediana –el estudiante que supera a la mitad de todos los estudiantes negros– calificaría aproximadamente en el percentil doceavo entre los estudiantes blancos. Una diferencia promedio de 0.8 desviaciones estándar (una de las diferencias de resultado más pequeñas que suelen encontrarse) apenas colocaría al estudiante afroamericano en la mediana en el percentil vigesimoprimero entre los estudiantes blancos. Es claro que son diferencias muy grandes, lo suficiente para tener implicaciones muy serias para el éxito posterior de los estudiantes.

Aunque todavía es muy grande, en las décadas recientes la brecha entre negros y blancos se ha reducido de manera considerable aunque errática. Prácticamente todos los datos confiables mostraron una lenta disminución en la brecha de las décadas de los sesenta y los setenta hasta los noventa o un poco después. En algunos periodos, esta disminución de la brecha reflejaba decrementos menores en los resultados de los afroamericanos que entre los blancos; otras veces, mostraba mayores ganancias. El progreso se detuvo alrededor de 1990, y en el curso de la siguiente década algunos datos indicaron que la brecha estaba aumentando de nuevo. Los datos de la NAEP indican que, en matemáticas, la disminución de la brecha se reanudó más o menos a finales de la década de los noventa en la escuela primaria y la secundaria, aunque el cambio entre los alumnos de 17 años era estadísticamente incierto. En el caso de la lectura, los datos de la NAEP

indican que la brecha entre los alumnos negros y blancos de primaria empezó a disminuir de nuevo más o menos al mismo tiempo, pero no hubo señal de mejora en el nivel de preparatoria y los datos sobre la secundaria (*middle school*) son contradictorios.

En promedio, también los estudiantes hispanos van muy a la zaga de los blancos no hispanos. En la actualidad, la diferencia promedio entre los hispanos y los blancos es en general más pequeña que entre los negros y los blancos, aunque en algunos casos es muy grande. Por ejemplo, la NAEP realizada en 2007 para matemáticas de octavo grado –que mostró una diferencia promedio de 0.90 desviaciones estándar entre negros y blancos– encontró una diferencia de 0.72 desviaciones estándar entre hispanos y blancos. No todas las pruebas presentan diferencias tan grandes entre hispanos y blancos, pero las diferencias considerables son la regla.

Sin embargo, es más difícil describir e interpretar la brecha entre blancos e hispanos. Los buenos datos acerca del desempeño de los estudiantes hispanos son más escasos y están disponibles por un lapso más corto que los datos sobre los estudiantes negros. La población hispana es sumamente diversa y algunos subgrupos, como los estadounidenses de origen cubano en Florida, muestran niveles relativamente altos de logro educativo, mientras que otros presentan niveles de logro mucho menores. Tal vez más importante es el hecho de que la población hispana se actualiza constantemente por la inmigración continua y los nuevos inmigrantes a menudo difieren considerablemente de quienes han estado por más tiempo en Estados Unidos. Por ejemplo, la elevada tasa de deserción entre los jóvenes hispanos ha sido foco de preocupación general durante décadas. Entre las décadas de los setenta y los ochenta no hubo mucho cambio consistente en la tasa de deserción de preparatoria de los estudiantes hispanos, pero esta estabilidad ocultaba dos tendencias compensatorias. Durante las primeras generaciones de residencia en Estados Unidos, la tasa de

deserción de preparatoria para los individuos de familias hispanas disminuyó gradualmente para acercarse a la norma de ese país. Sin embargo, ese progreso se vio oscurecido por la llegada continua de nuevos inmigrantes, muchos de los cuales no terminaban la preparatoria. De ahí que sea muy difícil esclarecer esas tendencias.

A pesar de los efectos de la rápida inmigración, los resultados de los estudiantes hispanos han mostrado algunas mejoras en relación con los de los estudiantes blancos no hispanos. Antes de 1980, diferentes bases de datos mostraban que los hispanos se acercaban a los blancos no hispanos, aunque esos incrementos eran modestos y en ocasiones inconstantes. Desde entonces la Evaluación Nacional ha proporcionado un cuadro contradictorio. En lectura, la evaluación de la tendencia a largo plazo de la NAEP, que tiene una muestra pequeña de hispanos y, por ende, es vulnerable a fluctuaciones que surgen del error de muestreo más que a verdaderos cambios en el logro, muestra pocos cambios significativos en la brecha entre hispanos y blancos en los grupos de menor edad, pero insinúa que la brecha entre los jóvenes de 17 años está empeorando. La NAEP principal, que es más grande, es más optimista y muestra una disminución modesta pero estadísticamente significativa de la brecha en cuarto grado entre 2000 y 2005 (se carece de datos de los alumnos de doceavo grado en 2005). Los resultados en matemáticas son parecidos, excepto en que muestran cierta evidencia de que la brecha también está disminuyendo en la secundaria.

No todas las diferencias raciales y étnicas favorecen a los blancos. En las pruebas de matemáticas y ciencia es común encontrar que los blancos van a la zaga de los estadounidenses de origen asiático. Por ejemplo, en la evaluación NAEP de matemáticas de doceavo grado realizada en el 2000, la calificación promedio de los estudiantes blancos estuvo 0.31 desviaciones estándar por debajo de la media del grupo de "asiáticos e isleños del Pacífico". Sin embargo, la interpretación de los resultados de los estudiantes de origen

asiático también es difícil por las mismas razones por las que es difícil interpretar las tendencias entre los hispanos. La categoría asiáticos/isleños del Pacífico incluye a un gran número de grupos que, fuera de la enorme porción del planeta de la que proceden, tienen relativamente poco en común y muestran patrones muy diferentes de desempeño educativo. La rápida inmigración ha convertido a este grupo étnico o racial en el de crecimiento más rápido en Estados Unidos; en términos de porcentaje, ese crecimiento es más rápido incluso que el de la población hispana y ha traído consigo cambios notables en la mezcla de grupos que comparten la etiqueta (por ejemplo, en la misma categoría se incluye a inmigrantes filipinos de bajos ingresos y a familias bien establecidas de origen chino y japonés). Además, a pesar de su rápido crecimiento, el tamaño de este grupo se mantiene relativamente pequeño en el país, por lo que los datos que describen su desempeño son limitados.

Es esencial tener en mente lo que esos datos nos dicen y lo que no. Sólo indican que las diferencias promedio entre los grupos raciales y étnicos son considerables. Pero aunque son muy grandes, esas diferencias no nos dicen nada sobre los estudiantes individuales. La variabilidad *en el interior* de cualquiera de esos grupos raciales o étnicos es muy grande, casi tanto como en la población estudiantil como un todo. Algunos estudiantes negros e hispanos obtienen resultados muy altos en relación con la distribución de los alumnos blancos, y hay estudiantes asiáticos cuyos resultados son muy inferiores al promedio de los blancos. Además, esos datos describen las diferencias entre grupos pero no las explican. La simple existencia de una diferencia grupal, incluso si es muy grande y repetida, no nos dice nada acerca de sus causas. Entre otros, los académicos han discutido por décadas acerca de las posibles causas, que incluyen escuelas de baja calidad, carencia de materiales, problemas de salud, bajo nivel educativo de los padres y diferencias culturales. Esa polémica es demasiado grande para abordarla aquí.

El desempeño en las pruebas y otras medidas de logro educativo también varían con la posición socioeconómica, un término algo amorfo acuñado por los sociólogos para describir la posición social de individuos y hogares. La posición socioeconómica suele malinterpretarse como sinónimo de ingreso, pero en realidad se refiere a algo más y a menudo se mide por un compuesto de ingreso, logro educativo y estatus ocupacional. En las descripciones de los resultados de las pruebas es frecuente que se calcule la posición socioeconómica (SES) con las medidas que se tengan a mano o puedan recabarse con facilidad, las cuales suelen ser débiles. Los estados y los distritos escolares sólo pueden recoger información limitada acerca de las características de los estudiantes (imagine que la directora de su escuela primaria le dice que tiene que proporcionarle información acerca de su ingreso anual) e incluso la mayoría de los estudios a gran escala que incluyen pruebas de logro, como NAEP y TIMSS, cuentan sólo con información insuficiente sobre la posición socioeconómica. Por ejemplo, NAEP no encuesta a los padres, lo que impide la obtención de información útil del ingreso familiar. Por lo tanto, frecuentemente tenemos que conformarnos con sustitutos débiles de los datos que en realidad queremos –por ejemplo, emplear la información de los estudiantes que califican para obtener almuerzos gratuitos o económicos como un sustituto de los índices de pobreza o ingreso y echar mano de las estimaciones que hacen los estudiantes del número de libros en sus hogares (¿qué tan bien podría usted calcularlo?) como un indicador de la formación educativa y la orientación de la familia.

Aunque estas mediciones endebles de la posición socioeconómica hacen que su relación con los resultados parezca más débil de lo que debería, por lo regular encontramos diferencias sorprendentes en el desempeño asociadas con la posición socioeconómica. Por ejemplo, en la evaluación de matemáticas de la NAEP para alumnos de doceavo grado realizada en el 2000, el desempeño

promedio de los alumnos que cumplían los requisitos para obtener almuerzos gratuitos o subsidiados estuvo 0.71 desviaciones estándar por debajo del resultado promedio de otros alumnos. Incluso las medidas más débiles de la posición socioeconómica, como las estimaciones que hacen los estudiantes del número de libros en sus hogares, por lo general predicen los resultados de las pruebas.

Comparaciones internacionales del desempeño de los estudiantes

Durante años, las comparaciones internacionales de los resultados de los estudiantes en pruebas de logro han aportado el combustible más potente para las críticas del sistema educativo estadounidense. Si bien las comparaciones internacionales son motivo de preocupación por el desempeño de los alumnos de ese país, el cuadro es menos sombrío y más complejo de lo que algunos críticos le han hecho creer.

Las comparaciones internacionales sistemáticas del logro de los estudiantes empezaron al menos desde 1959, con estudios *ad hoc* poco frecuentes que aplicaban pruebas comunes en los países que reunían apoyo para el esfuerzo. Desde mediados de la década de los noventa, esos empeños se institucionalizaron en tres series continuas de estudios. Una de ellas es el TIMSS, que en un principio se refería al Tercer Estudio Internacional de Matemáticas y Ciencia (*Third International Mathematics and Science Study*) pero que desde entonces recibió un nuevo nombre: Estudio Internacional de las Tendencias en Matemáticas y Ciencia (*Trends in International Mathematics and Science Study*), en reconocimiento al hecho de que se repite a intervalos de cuatro años. Las pruebas TIMSS se han empleado para evaluar a los alumnos hasta en cinco niveles de grado, aunque su mayor cobertura ha sido en octavo grado; esas

pruebas, diseñadas para reflejar elementos comunes del currículo, guardan un notable parecido con las pruebas de la Evaluación Nacional del Progreso Educativo. Aunque recibe un fuerte apoyo del Departamento de Educación de Estados Unidos, TIMSS opera bajo los auspicios de la Asociación Internacional para la Evaluación del Logro Educativo (*International Association for the Evaluation of Educational Achievement*, IEA), un consorcio internacional de organismos gubernamentales y organizaciones de investigación.

Un segundo esfuerzo, operado por el mismo grupo, es una evaluación de lectura conocida como el Estudio Internacional del Progreso en la Comprensión Lectora (*Progress in International Reading Literacy Study*, PIRLS). El tercer programa de evaluación, manejado por la Organización para la Cooperación y el Desarrollo Económicos (OCDE), es el Programa para la Evaluación Internacional de los Estudiantes (*Programme for International Student Assessment*, PISA). La prueba PISA, que se concentra sobre todo en matemáticas y lectura, tiene un propósito algo distinto: evaluar las competencias reales que han desarrollado los estudiantes para el momento en que se acercan a la terminación de la secundaria. El marco a partir del cual se elaboran las pruebas PISA no refleja muy de cerca los currículos escolares y las pruebas están organizadas en temas amplios, como "cambio y crecimiento", en lugar de áreas curriculares, como geometría.

Es frecuente que los resultados de esas evaluaciones internacionales se reporten como una sencilla clasificación de países y es común encontrar declaraciones sumarias como esta del Departamento de Educación de Estados Unidos: "En la [primera] evaluación TIMSS de octavo grado, los estudiantes de Estados Unidos obtuvieron resultados un poco por arriba del promedio internacional en ciencia y un poco por debajo del promedio en matemáticas".[5] Sin embargo, las comparaciones internacionales son un asunto complicado y no se justifica ese tipo de conclusiones simples.

La primera complicación es que las pruebas usadas para las comparaciones internacionales son, como todas las demás, pequeñas muestras de contenido, y las decisiones sobre el muestreo del contenido y el formato de su presentación son importantes. Por ejemplo, ¿qué porcentaje de los reactivos de la prueba de matemáticas debe asignarse al álgebra? (La decisión fue aproximadamente 25 por ciento en las pruebas TIMSS y NAEP pero sólo 11 por ciento en la prueba PISA.) ¿Qué aspectos del álgebra deberían enfatizarse y cómo deberían presentarse? ¿Qué formatos deben usarse y en qué combinación? TIMSS y NAEP usan el formato de opción múltiple en casi dos terceras partes de los reactivos mientras que PISA lo usa en una tercera parte.

Esas decisiones sobre la elaboración de las pruebas pueden tener un efecto considerable en los resultados incluso dentro de un país o un estado, pero son de particular importancia cuando se hacen comparaciones entre países debido a que las diferencias en sus currículos suelen ser grandes. PISA y TIMSS son pruebas muy diferentes que clasifican a los países de manera muy distinta. Por ejemplo, cuando se compararon los resultados de los 22 países participantes en ambas evaluaciones de matemáticas en secundaria, la clasificación de Escocia, Nueva Zelanda y Noruega fue mucho mejor en la evaluación PISA que en la de TIMSS; los Países Bajos, Hong Kong, Corea y Estados Unidos obtuvieron posiciones similares en ambas pruebas; y Rusia y Hungría se clasificaron mucho más alto en la evaluación TIMSS que en la prueba PISA.[6] En buena parte, esto es un reflejo de las diferencias en el contenido y las habilidades muestreadas por las dos pruebas. Incluso dentro de las restricciones de una prueba cualquiera, el muestreo importa. Por ejemplo, la clasificación de los promedios de los países puede ser modificada, aunque no mucho, cambiando simplemente el énfasis relativo que se da a las cinco áreas de contenido que componen la prueba de matemáticas de la TIMSS.[7]

Una segunda complicación es que las comparaciones internacionales no proporcionan un grupo normativo coherente y lógico para las comparaciones. En un reporte referido a normas, el grupo de comparación debe ser uno que dé a la comparación un significado útil –por ejemplo, reportar el desempeño de un estudiante como su rango percentil en una muestra de alumnos representativa nacionalmente, o al resultado promedio de una escuela en relación con la distribución de resultados de las escuelas del estado. Sin embargo, en los estudios internacionales el grupo de comparación es cuestión fortuita –incluye a cualquier país que haya aportado el dinero, el esfuerzo y el tiempo de clase para participar, y eso cambia de un estudio a otro, de un grado a otro y de un año a otro. Afirmaciones como "Estados Unidos obtuvo resultados en el promedio internacional" no significan mucho cuando ese promedio puede subir o bajar dependiendo de qué países sean elegidos para participar en una determinada evaluación. Esto no es sólo una posibilidad teórica. Por ejemplo, en el informe de la repetición de la TIMSS (conocida como TIMSS-R) en 1999, se informó que Estados Unidos había obtenido un resultado superior al "promedio internacional" de los países que participaron ese año. Unas páginas más adelante, el informe mostraba que el resultado obtenido por Estados Unidos era muy inferior al promedio de un grupo normativo diferente, los países que habían participado en 1995 y 1999.[8] Esta es una de las razones por las que la prueba PISA ofrece una perspectiva algo más pesimista del desempeño estadounidense que algunos de los resultados de la prueba TIMSS.

Esas complicaciones de ninguna manera hacen inútiles a las comparaciones internacionales, pero exigen una aproximación más prudente a la interpretación de los resultados. Las simples comparaciones con un "promedio internacional" no son de utilidad porque ese promedio es fortuito y algunas de las diferencias

son ambiguas porque una prueba diferente pero completamente razonable podría ocasionar un cambio en la clasificación.

¿Cómo puede entonces hacerse buen uso de las comparaciones internacionales? Primero: en lugar de comparar el desempeño con el promedio de los países que participaron, debemos comparar el desempeño de Estados Unidos con países específicos que, por alguna razón, son pertinentes. Por ejemplo, es útil comparar a Estados Unidos con países que en muchos sentidos son similares (digamos, Inglaterra y Australia) y con países con calificaciones particularmente elevadas a los que se desea emular. Si se procede de ese modo, los cambios casuales a lo largo del tiempo en los países participantes no tendrán efecto en nuestras conclusiones.

Segundo: no tratar a los resultados de cualesquier evaluación como algo más de lo que son: los hallazgos (algo proclives al error) de un muestreo particular del logro de los estudiantes. Esto tiene dos implicaciones. Las diferencias grandes y los patrones generales son más dignos de confianza que las diferencias pequeñas, de modo que debe prestarse poca atención a las últimas. Por ejemplo, la amplia brecha entre Estados Unidos y Japón en la más reciente evaluación TIMSS es lo bastante grande para ser considerablemente importante y sólida; cosa que no sucede con la minúscula diferencia entre Estados Unidos y Australia (el país clasificado justo por arriba de Estados Unidos).[9] Y estar alerta a las contradicciones entre evaluaciones –por ejemplo, entre los resultados de las pruebas PISA y TIMSS que pueden ser mayores de lo esperado. Dos o más fuentes de datos pintan un cuadro más completo y más confiable que sólo una.*

* Evaluaciones internacionales recientes proporcionan información sobre la significancia estadística para desalentar la atención de los lectores a las diferencias que son demasiado pequeñas para ser confiables, pero esta información refleja incertidumbre *dado el diseño de la prueba* y no toma en cuenta diferencias que podrían aparecer cuando se utilice una prueba con un diseño diferente.

Los países desarrollados del Oriente asiático de manera sistemática dominan el extremo superior de la distribución en las comparaciones internacionales del desempeño en matemáticas. Por ejemplo, en la evaluación TIMSS realizada en 2003, los estudiantes de Singapur, Corea, Hong Kong, Japón y Taipei obtuvieron los resultados promedio más elevados. El promedio de los estadounidenses cayó muy por debajo del de esos países asiáticos, pero fue muy similar a los promedios de muchas otras naciones que podrían considerarse competidoras. Por ejemplo, en la evaluación TIMSS que se hizo en 2003 de matemáticas de octavo grado, Rusia, Australia, Inglaterra, Escocia (examinada como un país separado en la prueba TIMSS), Nueva Zelanda e Israel obtuvieron resultados promedio que no eran significativamente diferentes a los de Estados Unidos. Los países europeos de mayor puntuación se ubicaron muy por encima de Estados Unidos, pero más próximos al resultado promedio de este país que a los de los países asiáticos. En el fondo de la distribución se encontraban sobre todo países menos desarrollados como Sudáfrica, Filipinas y Arabia Saudita, aunque hubo algunas excepciones como la de Chile.[10] Expresada en una forma estandarizada, la brecha entre el promedio de Estados Unidos y los promedios de los países con mejores promedios, como Japón y Corea, ha sido aproximadamente de 1 desviación estándar, aunque esto ha variado un poco de una iteración del estudio TIMSS a la siguiente.

PISA aplicó una prueba muy diferente a un grupo mayor de sustentantes, por lo que el cuadro resultante es muy diferente. En la aplicación de la prueba PISA en 2003, los resultados de Hong Kong, Corea y Japón se ubicaron de nuevo en el extremo superior de la distribución, pero los de Canadá y varios países europeos, como los Países Bajos y Bélgica, se ubicaron en el mismo rango general. En contraste también con los resultados de la prueba TIMSS, el desempeño de Australia y Nueva Zelanda fue significativamente mejor

que el de Estados Unidos en la evaluación de PISA. El resultado promedio de Estados Unidos se ubicó mucho más abajo en la distribución, de manera similar a los de Letonia y España.[11]

De este modo, tomados en conjunto, los dos programas internacionales de evaluación nos dejan con cierta incertidumbre acerca del desempeño relativo de los estudiantes estadounidenses en matemáticas. Puede confiarse en que el desempeño de esos alumnos es notablemente peor, en promedio, que el de los jóvenes de países desarrollados del Oriente asiático como Corea. Pero no queda clara la posición de Estados Unidos con respecto a los países europeos. Se han hecho muchas conjeturas acerca de las razones del desempeño inferior de los estadounidenses en la prueba PISA en comparación con la prueba TIMSS, pero a ese respecto sólo hay especulaciones. Las dos pruebas fueron diseñadas de forma tal que prácticamente no comparten contenido, lo que impide la realización de un análisis convincente de los efectos de las diferencias en la elaboración de las pruebas. Sólo podemos decir que los resultados dispares reflejan alguna combinación desconocida de las diferencias entre las pruebas y entre las muestras examinadas. Para explicar las diferencias entre sus hallazgos se necesitará algo más que una repetición de las pruebas de TIMSS y PISA, tal como se diseñan en la actualidad.

Aunque han atraído menos atención, las comparaciones internacionales del logro en lectura colocan a Estados Unidos en mejor posición. Un estudio internacional realizado hace una década y media, y que generó poco interés, demostró que el desempeño de los estudiantes estadounidenses en lectura era relativamente sólido en la secundaria y en el caso de la primaria ocupó el segundo lugar entre 28 países examinados.[12] Algunos escépticos sugirieron en ese momento que el estudio suscitó poca curiosidad porque las buenas noticias que ofrecía no eran de utilidad para quienes insistían en el fracaso de las escuelas estadounidenses. El Estudio Internacional

del Progreso en la Comprensión Lectora (PIRLS), que en muchos sentidos siguió el modelo de la prueba TIMSS, también muestra que el desempeño de Estados Unidos en lectura es bastante bueno en cuarto grado, el único grado examinado por la PIRLS. Estados Unidos se encontraba en un grupo de países con un elevado desempeño cuyos resultados promedio eran muy similares y que a nivel estadístico no diferían de manera significativa entre sí, como Canadá, Alemania e Italia. Sólo tres países –Suecia, los Países Bajos e Inglaterra– obtuvieron resultados promedio significativamente más altos que los de Estados Unidos. Los dos países desarrollados del Oriente asiático que participaron –Singapur y Hong Kong– obtuvieron calificaciones resultados promedio significativamente *menores*, pero eso puede reflejar las dificultades para aprender a leer en un idioma no alfabético.

Dado que nuestra atención debe enfocarse en el conocimiento y las habilidades que poseen los estudiantes cuando salen de la escuela, sería útil saber si la posición de los estudiantes estadounidenses en comparación con los de otros países mejora o empeora a medida que avanzan en la escuela. Muchos comentaristas han afirmado que empeora, pero la información que aborda esta cuestión es muy escasa. Varios de los estudios internacionales sólo examinaron una edad o grado escolar, y es arriesgado comparar diferentes pruebas aplicadas en edades distintas (como la de PISA y la TIMSS de la escuela secundaria) debido a que las discrepancias en los hallazgos pueden deberse a disparidades en las pruebas más que a diferencias en el desempeño relacionadas con la edad. La prueba TIMSS examinó a los alumnos en tres edades (en la escuela primaria, la secundaria y casi al final de la preparatoria), pero las muestras de los alumnos de preparatoria son tan diferentes entre los países que resulta difícil sacar conclusiones con alguna confianza. Esto nos deja con un puñado de comparaciones de estudiantes de primaria y secundaria.

Las mejores comparaciones son las proporcionadas por los datos de matemáticas de la prueba TIMSS que pertenecen a estudiantes de primaria y secundaria. Esos datos muestran que la brecha entre los alumnos de Estados Unidos y los de los países con mayores resultados tiende a ensancharse con la edad. Por ejemplo, la prueba TIMSS demuestra que la diferencia en el desempeño de Estados Unidos con Japón y Singapur es más grande en octavo que en cuarto grado. Esto apenas sorprende. Después de todo, la brecha aparece en cuarto grado porque en Singapur y Japón los estudiantes están aprendiendo a una tasa más rápida que los de Estados Unidos en los primeros grados, y no hay razón para esperar que su tasa de crecimiento disminuya al ritmo estadounidense una vez que ingresan al quinto grado. Esto no significa que el sistema de educación estadounidense sea peor, relativamente hablando, en los grados más altos. Si los factores que contribuyen al aprendizaje más rápido en Singapur y Japón (factores educativos como la calidad del currículo y factores no educativos como la cultura) se mantienen constantes entre los grados, cabría esperar un crecimiento sistemáticamente más rápido en Singapur y Japón, lo que ocasionaría que la brecha entre esos países y Estados Unidos aumentara de un grado al siguiente. Pero, en general, no sabemos lo bastante acerca del crecimiento relativo del desempeño de los estudiantes norteamericanos conforme avanzan en la escuela.

Por último, las comparaciones internacionales refutan una idea errónea común acerca de la variabilidad del desempeño de los estudiantes estadounidenses, aunque este hecho rara vez atrae la atención. Las diferencias penosamente grandes en el logro de grupos raciales/étnicos y socioeconómicos en Estados Unidos han llevado a mucha gente a suponer que los estudiantes de este país deben variar más que otros en el desempeño educativo. Algunos observadores incluso han dicho que por este motivo la competencia del tipo de carreras de caballos (con comparaciones simples de

los resultados promedio entre países) es engañosa. Los estudios internacionales abordan esta cuestión, aunque con una advertencia: la estimación de la variabilidad en los estudios internacionales es mucho más débil que la estimación de promedios. Esto se debe tanto al diseño de las muestras como a aspectos misteriosos de los modelos matemáticos usados para poner los resultados en una escala. Por lo tanto, no tiene sentido ofrecer conclusiones como "la desviación estándar del desempeño en Japón es 10 por ciento más grande que la de Estados Unidos". Estamos limitados a conclusiones más generales, algo parecido a "las desviaciones estándar de Estados Unidos y Japón son muy similares".

Cosa que así es. En efecto, la variabilidad del desempeño de los estudiantes es muy similar en la mayoría de los países, independientemente del tamaño, la cultura, el desarrollo económico y el desempeño del estudiante promedio. Por ejemplo, 31 de los países que participaron en la primera evaluación TIMSS de matemáticas de octavo grado cumplieron las normas de calidad de datos para el estudio. De ellos, todos salvo cuatro tenían una desviación estándar de calificaciones en el rango de 73 a 102 puntos. Contrario a la especulación de que Estados Unidos mostraría una variabilidad particularmente grande debido a su heterogeneidad social, la variabilidad de ese país estaba muy cerca de la mitad del grupo, con una desviación estándar de 91. De igual modo, en la evaluación PISA de matemáticas aplicada en 2003, la variabilidad del desempeño de los estudiantes estadounidenses estuvo muy cercana al promedio de 29 países de la OCDE.

Todavía más sorprendente fue el ordenamiento de los países: en general, la homogeneidad social y educativa de los países no predice homogeneidad en el desempeño de los estudiantes. En algunos países pequeños y homogéneos (como Túnez y Noruega) las desviaciones estándar de los resultados son relativamente pequeñas, pero en muchos otros no. Por ejemplo, en los tres estudios

TIMSS, la desviación estándar de los resultados en Japón y Corea (dos países con mayor homogeneidad social que Estados Unidos y en los cuales los sistemas de educación son más homogéneos hasta el octavo grado) era aproximadamente similar a la desviación estándar de Estados Unidos. La evaluación PISA también confirmó este hallazgo general, demostrando que la variabilidad en el desempeño en Estados Unidos es un poco mayor que en Corea pero más pequeña que en Japón. Los resultados de las evaluaciones TIMSS y PISA también demostraron que la relación que se encuentra en Estados Unidos entre los resultados y los antecedentes socioeconómicos es generalizada. Aunque hay algunas excepciones (por ejemplo, Macao e Islandia en la evaluación PISA de 2003), en la mayor parte de los países apareció una relación considerable entre los resultados y la posición socioeconómica (SES) en ambas evaluaciones, a pesar de que en ambas la medición de la posición socioeconómica era débil.

Esto no quiere decir que la enorme variabilidad que vemos en el desempeño de los alumnos en las pruebas sea inalterable. Sin embargo, esos datos muestran que las bases de esta variabilidad son complejas y que pueden esperarse grandes variaciones incluso en los sistemas educativos de alto desempeño y más equitativos. Esto, como veremos, es un problema cuando se establecen expectativas para la reforma educativa en Estados Unidos. ■

El ABC
de la
evaluación educativa

Capítulo 6
¿Qué influye en los resultados de las pruebas? (o cómo no escoger una escuela)

En 1990 se dieron a conocer los primeros resultados de la NAEP que clasificaban a los estados, en forma de ranking, en función de sus resultados promedio. En la parte superior de la distribución se encontró una representación excesiva de estados de la región norcentral y de Nueva Inglaterra, entre los que se ubicaban Minnesota, Maine, Vermont, Massachusetts y Dakota del Norte. Al fondo se agruparon de manera desproporcionada estados del sureste como Mississippi, Alabama, Arkansas y Louisiana. Aunque hubo algunos resultados inesperados, el patrón regional no fue una sorpresa ya que durante mucho tiempo se había observado (dicho patrón) en otros datos menos accesibles.

Esos resultados suscitaron mucha discusión. Parecía que casi todos los que hablaban en público (comisionados estatales de educación, reformadores y críticos de la educación, periodistas) tenían a la mano una explicación de la posición de los estados que les interesaban, la cual por lo general daba crédito o descrédito a algún aspecto de los sistemas educativos estatales. Por ejemplo, un comisionado se apresuró a anunciar que sería necesario sustituir el currículo de matemáticas de su estado. Casi perdido entre la multitud se encontraba el comisionado de un estado de la región norcentral cuyos resultados se encontraban entre los más altos, como suele suceder con los estados de esa zona. Cuando se le preguntó a qué se debía el excelente desempeño de sus estudiantes, respondió que en su estado no había playas o montañas que los distrajeran.

¿Hablaba en serio o sólo se estaba burlando de las necedades de quienes lo rodeaban? Nunca tuve la oportunidad de preguntárselo.

No sorprende que todos esos comentaristas se aferraran a las supuestas explicaciones de las diferencias en el desempeño de los estados. Después de todo, a pocos nos interesan las descripciones del logro por sí mismas. La mayoría de la gente quiere saber qué resultados son correctos, cuáles no lo son, a quién o a qué reconocer o culpar y cómo arreglar lo que no funciona. Para mejorar el sistema educativo se requiere identificar los programas, escuelas y sistemas eficientes e ineficientes, de modo tal que puedan emularse los primeros y ponerse fin a los segundos. Si se va a seguir recompensando y castigando a los educadores por los resultados obtenidos en las pruebas, es menester identificar las escuelas pertinentes para ambas cosas.

No obstante, sus aseveraciones fueron, en su mayor parte, desatinadas. Los resultados obtenidos en las pruebas describen lo que los estudiantes saben y pueden (o no pueden) hacer. En la mayoría de los casos, eso es lo que hacen *todos* los resultados de las pruebas. Salvo en circunstancias inusuales, como en un experimento planeado con asignación aleatoria de los participantes, los resultados por sí mismos no explican *por qué* los estudiantes saben lo que saben. Para eso se necesitan datos adicionales, a los cuales rara vez tienen acceso los comentaristas que suelen apresurarse a informar por qué los alumnos pueden o no pueden desempeñarse como se desea. Es posible que algunos de los comentarios sobre los resultados de la NAEP fueran correctos, pero cuando eso sucedió por lo general se debió al azar.

La razón de que sus afirmaciones por lo regular fueran infundadas es sencilla: en el logro educativo influyen muchas otras cosas además de la calidad de las escuelas, y el impacto de esos factores ajenos a la educación puede ser enorme. Cuando una escuela tiene un desempeño bueno o malo en una prueba de logro, la razón

puede ser la calidad de la educación, cualquier cantidad de causas no educativas o –lo que es más probable– ambas cosas. No siempre es fácil averiguar cuál es el caso. Esos comentaristas y políticos se parecen a la compradora frustrada de una casa que describí en el capítulo 1: una persona ansiosa por inferir la calidad de una escuela a partir únicamente de los resultados de las pruebas, sin querer hacer el trabajo duro de buscar los datos adicionales que se necesitarían para identificar las diferencias en la eficacia educativa.

Esto no es sólo una preocupación académica. De manera rutinaria la gente malinterpreta las diferencias en los resultados obtenidos en las pruebas y suele dar más peso del adecuado a la calidad de la educación. Es común que en las tendencias de los resultados a lo largo del tiempo (sean descendentes o ascendentes) influyan factores sociales y, en el caso de mejoras aparentes, una enseñanza inapropiada para la prueba. No todas las escuelas que obtienen bajos resultados ofrecen un programa educativo tan deficiente como podrían sugerir sus resultados. Por la misma razón, si las escuelas de su vecindario obtienen altos resultados, eso podría indicar menos acerca de la calidad de sus programas de lo que a usted le gustaría. La enorme variabilidad en los resultados mostrada por los estudiantes estadounidenses no es anómala y sus causas son distintas a lo que suele pensarse. Entender esas complejidades no sólo ayuda a los padres a hacer una elección atinada de las escuelas, también nos permite evaluar el éxito o el fracaso de las políticas educativas y establecer objetivos razonables para las reformas.

Cualquiera que estudie el logro educativo sabe que las diferencias en los resultados de las pruebas se deben en buena medida a factores no educativos. Eso ha sido documentado por una enorme cantidad de investigaciones realizadas a lo largo de medio siglo en Estados Unidos, país que no se sale de lo común en este aspecto. En la actualidad se libra un acalorado debate acerca del impacto *relativo* de los factores educativos y los no educativos. Buena parte de la

investigación sugiere que las variaciones en los factores sociales explican la mayor parte de la variabilidad de los resultados, aunque algunos investigadores sostienen que ello refleja defectos en el diseño de los estudios (como la medición inadecuada de la calidad educativa) y que algunos factores educativos (en particular, las variaciones en la calidad de los maestros) tienen gran influencia en los resultados.* Por el momento no es posible zanjar esta discusión debido en parte a que las diferencias en la calidad educativa están fuertemente confundidas con los factores sociales que influyen en los resultados. En promedio, la calidad de los recursos educativos es inferior en los barrios pobres con una elevada proporción de grupos minoritarios que en los vecindarios ricos, y es extremadamente difícil desentrañar los efectos de los factores educativos y los sociales. No obstante, es indiscutible que los factores sociales tienen un impacto considerable en los resultados, incluso si no puede establecerse de manera precisa la proporción de la variabilidad que explican.

El impacto de los factores no educativos en los resultados obtenidos en las pruebas es un ejemplo de lo que uno de mis compañeros de posgrado llamaba un "hallazgo de la abuela", algo de lo que uno podría haberse enterado preguntando a cualquier abuela razonablemente sagaz sin tener que tomarse la molestia de conducir

* Esta controversia está ligada a otra que ha recibido mucha atención de los políticos y de la prensa: si debería evaluarse a las escuelas con base en los cambios de resultados de una cohorte a otra (por ejemplo, comparar a los alumnos de cuarto grado de este año con los del año pasado, tal como lo exige la ley NCLB) o por medio del llamado análisis de valor agregado que sigue a los estudiantes y evalúa cuánto han mejorado en el curso del año escolar. Varios estudios de valor agregado pretenden mostrar efectos muy grandes de las variaciones en la calidad de los maestros. Una revisión reciente de esos estudios, en la que participé, llegó a una conclusión más cauta: las investigaciones en realidad mostraban un efecto importante de la calidad de los maestros pero los resultados no eran todavía suficientes para establecer bien su tamaño. Véase D. F. McCaffrey *et al.*, *Evaluating Value-Added Models for Teacher Accountability*, MG-158-EDU (Santa Monica, CA: Rand, 2003).

estudios científicos. Por ejemplo, imagine dos escuelas hipotéticas. La escuela A tiene una población estudiantil estable, compuesta en su totalidad por hablantes nativos del inglés de origen moderadamente privilegiado. En contraste, muchos de los alumnos de la escuela B son inmigrantes, la mayoría con un dominio limitado del inglés y una tasa muy alta de rotación, por lo que los profesores tienen una oportunidad limitada de trabajar con algunos de los estudiantes antes de que se vayan y sean reemplazados por otros recién llegados. Suponga ahora que esas dos escuelas obtienen la misma puntuación promedio o, en la métrica del programa NCLB, el mismo porcentaje de alumnos por arriba del estándar de competencia. En términos de la mayor parte de los sistemas actuales de rendición de cuentas en la educación, esas dos escuelas reciben un trato idéntico. Pero suponer que la calidad de la educación es similar en ambas escuelas desafía la lógica. Los profesores de la escuela A tienen un trabajo mucho más sencillo que los de la escuela B, y si su trabajo fuera tan eficiente deberían haber producido puntuaciones promedio mucho más elevadas. Dada la información adicional que proporcioné acerca de los estudiantes de ambas escuelas, parece probable que la calidad de la educación sea menor en la escuela A que en la escuela B, pero las puntuaciones no habrían revelado esto por sí solas.

A pesar de ser tan evidente, por lo general se ignora el impacto de los factores no educativos en los resultados; es algo que hacen los políticos que desean reclamar el crédito o tener a quien culpar; la prensa, que quiere publicar un reportaje que llame la atención; los consumidores y los agentes del mercado inmobiliario; y también, con demasiada frecuencia, los educadores. Algunas personas parecen no estar enteradas; otras entienden que las diferencias no educativas son importantes, pero no lo suficiente para preocuparse. Es común escuchar frases como "lo perfecto es enemigo de lo bueno", lo que implica que preocuparse en exceso por

la exactitud de los datos (o, para ser más precisos, por la exactitud de la inferencia basada en los datos) es un obstáculo para la obtención de información útil.

En 1984, el Departamento de Educación de Estados Unidos publicó la primera de las llamadas Listas de Ranking en que se comparaba a los estados en términos de los resultados promedio obtenidos en la prueba SAT. La intención era identificar a los estados con mejores sistemas educativos. Muchos alegamos que había muchas razones por las que dicha comparación era esencialmente engañosa, siendo la más importante el hecho de que la proporción de estudiantes que presentaban la prueba variaba mucho entre los estados. Los estados en que grupos pequeños y muy selectivos presentaban la prueba (por ejemplo, en esos días en algunos estados de la región central muy pocos estudiantes presentaban la prueba SAT, en su mayoría eran personas con un logro muy alto que estaban interesadas en ingresar a universidades competitivas de la costa este u oeste) ascendían en la clasificación debido simplemente a las características de los sustentantes. Atribuir esas diferencias en los resultados promedio a supuestas diferencias en la calidad educativa era un sin sentido. Nadie podía ofrecer una refutación de peso; este problema era bien conocido y bastaba con la simple aritmética para demostrar su gravedad. Sin embargo, la respuesta ofrecida fue que lo perfecto era enemigo de lo bueno.

La idea de que "lo perfecto es enemigo de lo bueno" sería una orientación razonable si los datos defectuosos proporcionaran una respuesta aproximadamente correcta. Pero resultaría un pésimo consejo si las respuestas sugeridas por dichos datos fuesen básicamente erróneas, como fue el caso de las listas de ranking del Departamento de Educación y, en menor medida, de las comparaciones simples de los promedios estatales en la NAEP. Es común que las personas no especializadas y los expertos discrepen acerca de dónde cae la línea entre los resultados que son útiles pero no

precisos y aquellos que están demasiado alejados de ser engañosos. Esto puede ayudar a explicar la propensión a ignorar las influencias no educativas en los resultados. Hace algunos años me pidieron que asistiera a una reunión en un periódico para discutir la costumbre del diario de reportar los resultados de las pruebas estandarizadas del estado, reporte que incluía las tablas que mostraban los porcentajes de alumnos en los distritos y preparatorias cuyos resultados eran iguales o superiores a ciertos estándares de desempeño. Muchos de los miembros del personal insistían en saber si estaban usando la métrica correcta, pero algunos de nosotros, todos científicos sociales, recalcábamos que el problema era más grave: la inferencia que muchos lectores podrían extraer de esas tablas (que las escuelas con mayores resultados eran mejores) era injustificada y, en algunos casos, terriblemente engañosa. Hicimos pocos progresos. No creo que ninguno de los presentes estuviera dispuesto a aceptar resultados erróneos; me parece que los investigadores no logramos persuadirlos de que los resultados, tal como se presentaban, podían estar equivocados. Los otros se encontraban todavía en el lado de "lo perfecto es enemigo de lo bueno".

Algunos críticos de la educación hacen caso omiso de las explicaciones no educativas de los resultados obtenidos en las pruebas, no porque tengan evidencia verosímil de lo contrario sino porque creen que dichas explicaciones eximen a las escuelas de su responsabilidad. Muchos críticos conservadores no quieren que se oculten los defectos del sistema educativo como un todo, mientras que algunos críticos liberales temen que el reconocimiento de esas influencias disculpará el bajo desempeño de las escuelas que atienden a estudiantes de bajos ingresos y de grupos minoritarios. Para ser justos, algunos de esos críticos en realidad no niegan el impacto de los factores no educativos sino que, más bien, fingen no verlos por temor a desviar la atención de las variaciones en la calidad educativa que importa.

Lo irónico es que esos críticos suelen perder de vista el otro lado de la moneda: al ignorar las causas no educativas de las variaciones en los resultados (es decir, al asumir que los resultados son una indicación directa de la calidad de la escuela) se permite que algunas escuelas con *altas puntuaciones* queden eximidas de su responsabilidad. Los altos resultados obtenidos por algunas escuelas no se deben tanto a la calidad de la educación que ofrecen, sino a los alumnos a los que atienden; pero quienes están convencidos de que los resultados reflejan sobre todo la calidad educativa las consideran buenas escuelas. Mis hijos asistieron a algunas de las escuelas con mayores puntuaciones en nuestro estado. Algunos de sus maestros eran verdaderamente magníficos, pero también tuvieron algunos profesores mediocres y unos cuantos a los que en mi opinión no se les debería haber permitido enseñar, como un profesor de inglés que en el día de visita de los padres cometió errores gramaticales y de vocabulario tan atroces que desató protestas repetidas y audibles entre los padres sentados en la parte trasera del salón. Sin embargo, las puntuaciones siempre eran altas, lo que en parte era un reflejo del muy elevado nivel educativo de la comunidad que estaba llena de abogados, médicos, académicos, economistas, diplomáticos, investigadores en biomedicina y cosas por el estilo. Otro hallazgo de la abuela. Esos padres no sólo proporcionaban, en promedio, ambientes muy propicios para el logro académico, sino que muchos brindaban también instrucción complementaria, fuese reenseñando directamente el material o pagando por los servicios de empresas de tutoría en el vecindario.

Un ejemplo concreto: cuando mi hijo estaba en séptimo grado, llevaba una clase de matemáticas que no era bien enseñada (fui a observarla para confirmar mi corazonada). Una noche me dijo que no entendía su tarea de matemáticas, que era parte de una introducción a la probabilidad y la estadística. Primero traté de aclarar la tarea, pero muy pronto me di cuenta de que le faltaban algunas

nociones clave. Le pedí los materiales de su clase, los revisé, volví a enseñarle algunos de los conceptos fundamentales, después de lo cual pudo encargarse de la tarea. Regresé a la cocina para lavar los platos de la cena, pero pronto volvió a llamarme. Acababa de hacer una audición exitosa para la banda de jazz de la escuela, pero le resultaba difícil contar los ritmos en la pieza que tenía que practicar. Los conté por él, pero todavía le resultaba confuso (tal como me había sucedido a mí, muchos años antes, cuando empecé a tocar jazz). De modo que fui a buscar mi trompeta y toqué la música a medio tempo mientras él contaba. Eso funcionó. Cuando regresé a fregar cacharros, mi esposa volteó y me dijo, "Ahí lo tienes: diferencias de clase social en el logro educativo".

Los críticos que ignoran el impacto de los factores sociales sobre los resultados no captan la idea: la razón para reconocer su influencia no es dejar que alguien quede exento de su responsabilidad sino obtener la respuesta correcta. Ciertamente, las bajas puntuaciones son una señal de que algo anda mal; después de todo, una de las razones principales para aplicar las pruebas es averiguar los puntos en que el desempeño es sólido o débil. Pero de la misma manera en que la fiebre no puede revelar por sí misma cuál es la enfermedad de un niño, los bajos resultados no dicen por sí mismos *por qué* es bajo el desempeño y por lo regular son insuficientes para revelar dónde es buena o mala la enseñanza. Los resultados decepcionantes pueden enmascarar una buena enseñanza mientras que las altas puntuaciones pueden ocultar problemas que es necesario abordar. Esto sucede sobre todo ahora que la evaluación de alto impacto puede dar lugar a una grave inflación de los resultados, como se explica en el capítulo 10. Los bajos resultados, como la fiebre, son una indicación de que se necesita una mayor exploración para averiguar lo que en realidad está mal y las intervenciones que podrían ser de utilidad para enfrentar el problema. Si se utilizan de manera apropiada, los

resultados pueden ofrecer información descriptiva inestimable que puede usarse para mejorar la educación, pero de nada sirve hacer una evaluación incorrecta de las escuelas o los sistemas educativos que atribuya el crédito o la culpa a quien no corresponde.

Identificación de las causas de los resultados

Ni mi esposa ni yo somos médicos, de modo que cuando uno de nuestros hijos presentaba fiebre por lo general no sabíamos a qué obedecía. La temperatura del niño era un indicador de que estaba enfermo, pero no era suficiente para revelar la causa. En ocasiones teníamos un presentimiento, pero rara vez era lo bastante bueno para que estuviésemos dispuestos a actuar en función del mismo. Después de todo, muchas enfermedades pueden ocasionar una fiebre repentina, de modo que llamábamos al pediatra. Él nos solicitaba la información descriptiva que habíamos reunido, como qué tan alta era la fiebre, y por lo regular nos preguntaba otras cosas, como detalles de los síntomas gastrointestinales del niño, si había tenido dolor de cabeza y cosas por el estilo. Luego aventuraba su hipótesis acerca de la causa de los síntomas o nos pedía que lleváramos al niño para poder obtener más datos. A menudo su respuesta no era definitiva, sino más bien en la línea de "Esto es suficiente para descartar X y Y, que eran las posibilidades preocupantes; lo más probable es que se trate de Z, vamos a dar tratamiento para eso y veamos qué sucede en las próximas 24 horas".

Con frecuencia pasamos por un proceso similar cuando nuestro automóvil falla. Cuando yo era más joven, los carros eran más simples y yo parecía tener más tiempo, afinaba y a veces reparaba mi carro por mí mismo. En estos días, dada la complejidad y los sistemas de control por computadora de los carros modernos,

llevo el carro al mecánico. Aunque su proceso de diagnóstico es incomparablemente más experto, en esencia es igual al que yo utilizaba y es el mismo que empleaba el pediatra, sugerir posibles explicaciones para el problema observado mediante la obtención de información descriptiva, remitirla al conocimiento experto y descartar las explicaciones alternativas.

Usted puede encontrar ejemplos maravillosos de este tipo de razonamiento si escucha a Click y Clack, los tipos del programa *Car Talk* en la Radio Pública Nacional.

De modo que la mayoría de nosotros estamos más o menos familiarizados con el proceso de pasar de los datos descriptivos a las explicaciones y nos damos cuenta de que esto suele ser un trabajo difícil. Pero dado que este proceso común resulta ser demasiado polémico cuando los datos que hay que explicar son los resultados de las pruebas, vale la pena dedicar un poco más de tiempo a poner de relieve algunos de los principios formales que rigen este proceso científico.

En los ejemplos comunes presentados antes se oculta un principio que es axiomático para los científicos, aunque para los demás suele ser confuso. La mayoría de la gente piensa que una hipótesis se confirma cuando los datos reunidos son congruentes con ella. No obstante, eso es sólo parte del proceso. *Una hipótesis sólo tiene credibilidad científica cuando la evidencia reunida ha descartado explicaciones alternativas plausibles*. Esto es fundamental para la ciencia y usted puede escucharlo cada semana en las discusiones entre Click y Clack (ambos graduados del MIT), en afirmaciones como "No, no pueden ser los cojinetes de la rueda porque el ruido desaparece cuando se aprieta el embrague".

La incapacidad para reconocer este principio es tal vez la razón principal por la que mucha gente supone que los resultados de las pruebas son por sí mismos suficientes para indicar la eficacia o calidad de una escuela. A las personas que usaban los resultados

como argumento de que las escuelas de mis hijos eran de alta calidad, un científico social competente les habría respondido: "Puede ser, pero ¿tuvieron en consideración los *otros* factores que podrían dar una explicación convincente de esas altas puntuaciones, como el nivel educativo de los padres de la comunidad?".

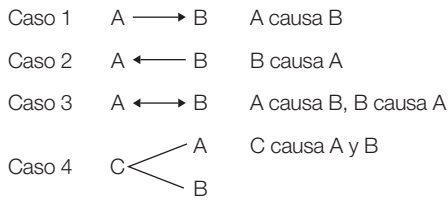
La pregunta no implica que se tenga que elegir una u otra opción: tanto la calidad de la educación como otros factores no educativos (la educación de los padres, por ejemplo) pueden haber contribuido de manera conjunta a los altos resultados obtenidos en las escuelas de mis hijos. (Creo que un estudio cuidadoso, que no se ha realizado, habría revelado precisamente eso.) El punto es que no puede asignarse todo el crédito ni toda la culpa a un factor, como la calidad de la escuela, sin investigar el impacto de los otros. Muchos de los modelos estadísticos complejos que se emplean en economía, sociología, epidemiología y otras ciencias son esfuerzos por tomar en cuenta (o "controlar") otros factores que ofrecen explicaciones alternativas verosímiles de los datos observados, y muchos distribuyen la variación en el resultado (digamos, los resultados obtenidos de las pruebas) entre varias causas posibles. En muchos tipos de investigación se considera que los experimentos verdaderamente aleatorizados son la regla dorada porque eliminan la mayor parte de las explicaciones alternativas (porque los grupos comparados son equivalentes salvo en el tratamiento aplicado).

Aunque casi todos hemos pasado por este proceso de respaldar la explicación propuesta por una hipótesis descartando alternativas plausibles, esa no parece ser la forma en que la mayoría de la gente piensa respecto a los resultados de las pruebas obtenidos en la educación. Más bien, lo común es que adviertan que uno o dos factores concuerdan con algún patrón de los resultados y que luego anuncien que estos factores son la explicación del patrón. Lo más frecuente es que los factores que reciben el crédito o el descrédito sean aspectos de la educación.

Eso nos lleva directamente a un segundo principio para el descubrimiento de causas: *una simple correlación no necesariamente indica que uno de los factores es la causa del otro*. Por ejemplo, considere el siguiente hallazgo de la Evaluación Nacional del Progreso Educativo: "En el año 2000, los maestros de octavo grado que informaron que habían asignado tareas que se llevaban 45 minutos tenían estudiantes con calificaciones promedio más altas que los alumnos de profesores que les asignaban menos tarea".[1] Esta es una correlación positiva, un término que tiene significado técnico pero que en el lenguaje no especializado significa que el aumento en un factor (la tarea) se asocia con el aumento en otro (las calificaciones). (Una correlación negativa significa que el aumento de un factor se relaciona con la disminución en el otro –por ejemplo, si más tarea se asociara con menores calificaciones– y una correlación cero significa que no existe relación entre ambos factores.) La asociación positiva entre la tarea y las calificaciones es un hallazgo común, y una interpretación frecuente es que la tarea ocasiona un incremento en las calificaciones. Quizá, pero quizá no.

Existen otras explicaciones posibles para la relación observada. La figura 6.1 muestra cuatro casos generales en que A es una variable (que en este ejemplo puede ser la tarea) y B la otra (en este caso, las calificaciones obtenidas). Más tarea puede ser la causa de calificaciones más altas (caso 1). Una alternativa sería que lo contrario fuese la verdad (caso 2): un mayor logro puede dar lugar a más

Figura 6.1. Cuatro explicaciones causales de una correlación simple



tarea. Al principio esta posibilidad puede parecer un poco forzada, pero en realidad no es así: es posible que los maestros a los que se les asignan chicos de alto rendimiento piensen que sus alumnos pueden hacer más y beneficiarse de ello, por lo que les encargan más tarea. El caso 3 sólo dice que el proceso causal puede operar en ambas direcciones a la vez.

El caso 4 atribuye por completo a otro factor el crédito o el descrédito: otra cosa (el factor C) es la causa del aumento en las tareas y en las calificaciones. Por ejemplo, es posible que no sea la tarea la que ocasione el incremento en las notas sino, más bien, que algunos atributos de ciertos padres (por ejemplo, un mayor interés por el logro académico) lleven a sus hijos a obtener calificaciones más altas y a sus maestros (preocupados por recibir la aprobación de los padres) a asignarles más tarea. (Este proceso fue seguramente el que operó en la secundaria de mis hijos. Un año en que me quejé de que la tarea era excesiva, el director respondió que algunos padres se quejarían si se encargara menos tarea.) Dick Darlington, profesor en Cornell, solía presentar el siguiente ejemplo en sus clases de estadística: la cantidad de sal que se vierte en las carreteras se correlaciona con la tasa de accidentes automovilísticos. ¿La sal (digamos el factor A en la figura 6.1) ocasiona los accidentes (el factor B en la figura 6.1)? Por supuesto que no. El elemento que falta es el factor C: la nieve. Cuando hay nieve en los caminos, el personal de carreteras vierte sal, pero esto es insuficiente para compensar por completo lo resbaladizo del camino ocasionado por el clima, y se incrementa la tasa de accidentes.

El punto no es argumentar en favor de una de esas explicaciones, sino el hecho de que existen muchas posibilidades y que *sin información adicional no podemos saber cuál es la correcta*. La gente elige a menudo la explicación que parece más razonable. Por desgracia, no siempre hay acuerdo respecto a qué es más razonable y, en cualquier caso, sus conjeturas suelen ser erróneas.

Es muy fácil ignorar este principio cuando una explicación causal parece de sentido común (¿no debería mejorar el aprovechamiento con más tarea?), pero resulta más claro cuando se considera una mayor variedad de correlaciones. Hace muchos años empecé a escribir una broma, basada en hallazgos reales de la investigación, titulada "Los efectos de los trabajos irrelevantes del curso". Las investigaciones publicadas muestran muchas correlaciones entre el trabajo en el curso y las calificaciones. Por ejemplo, más trabajo en matemáticas se asocia con mejores calificaciones en esa materia. Este tipo de hallazgo suele ser anunciado en los medios con bombo y platillo y recibe mucha atención de la comunidad. La interpretación es congruente: los niños deberían llevar más cursos de la materia X para mejorar su desempeño en esa materia. Hasta cierto punto, esa conclusión es obviamente razonable: la razón de llevar un curso de álgebra es aprender álgebra, y los alumnos que más la estudian aprenderán más. Sin embargo, no acaba aquí la historia. La complicación –suele ignorársele en los reportes de prensa– es que más trabajo en algunas materias predice también mayores resultados en *otras* materias. Por ejemplo, hace algunos años, H. D. Hoover, en ese entonces director de los Programas de Evaluación de Iowa, demostró que los alumnos de preparatoria que llevaban más cursos de matemáticas obtenían resultados más elevados no sólo en la prueba de matemáticas de la ACT, sino también en una combinación de las otras tres pruebas de la ACT (inglés, ciencia y estudios sociales). El mismo patrón se encuentra en los cursos de ciencia.[2] Otras investigaciones encontraron que el estudio de lenguas extranjeras predice los resultados en lectura y vocabulario de inglés. Podría concluirse que los estudiantes aprenden el significado de las raíces de las palabras cuando estudian las lenguas romances, pero espere: también predice los resultados en matemáticas.[3] ¿A qué se debe eso? ¿El aprendizaje de las conjugaciones francesas de verdad facilita el aprendizaje del álgebra? ¿Llevar más cursos

de ciencia mejora las habilidades de lectura y el conocimiento de los estudios sociales? Tal vez un poco, pero con certeza no mucho. La clave está en la tercera variable: las características de los chicos que llevan muchos cursos de matemáticas y que estudian desde temprano un idioma extranjero. Esos muchachos suelen tener una elevada motivación de logro y, por lo general, su desempeño es alto. Un colega lo decía de manera sencilla en términos políticamente incorrectos: "A los chicos listos les va bien en las pruebas".

El hecho de no tomar en cuenta este factor particular (es decir, las características de la gente con valores más altos en una de las variables) es un problema clásico de la ciencia social que lleva la etiqueta formal de *sesgo de selectividad*. Considere el siguiente ejemplo: hace algunos años, cuando una gran proporción de estudiantes no llevaban cursos de álgebra, se publicó un estudio que demostraba que los alumnos que llevaban álgebra tenían mayor probabilidad de concluir la universidad que quienes no lo hacían. En la comunidad de la política educativa por lo general se interpretaba que esto significaba que A causa B: que llevar álgebra *hace* más probable que los estudiantes concluyan la universidad. Se llegó a conocer al álgebra como el curso "filtro" y se hicieron grandes esfuerzos por aumentar la proporción de estudiantes que lo llevaban. En mi opinión, este esfuerzo fue para bien: aprender algo de álgebra es bueno, puede ayudar a la gente a entender todo tipo de problemas en la vida real y ayuda a abrir puertas para los estudiantes en sus estudios posteriores y en el mundo del trabajo. Sin embargo, la interpretación de este estudio particular era sospechosa por el sesgo de selectividad: los estudiantes que en esos días llevaban cursos de álgebra eran diferentes a los que no lo hacían, y algunas de las diferencias entre los grupos pueden haber contribuido a la correlación observada. Por ejemplo, los estudiantes que querían ingresar a universidades competitivas (en su mayoría alumnos de alto rendimiento) por lo general llevaban álgebra.

Como empieza a apreciarse en esos ejemplos, cuando queremos explicar los resultados obtenidos en las pruebas pueden omitirse terceros factores (como el que lleva la etiqueta C en la figura 6.1) que presentan una variedad de explicaciones plausibles de los resultados que es necesario refutar. La lista de factores que se ha demostrado que predicen el logro, incluyendo los resultados de pruebas, es larga y diversa. Incluye, por ejemplo, la salud de los niños que presentan las pruebas, la educación de los padres, el ingreso, la configuración de la familia y las influencias culturales. La lista es demasiado grande y la controversia sobre la evidencia es demasiado extensa para abordarla aquí. Pero usted podría preguntarse: ¿Cómo nos va en el intento de "controlar" o ajustar los efectos de esos diversos factores para aislar lo que la gente en realidad quiere medir: el efecto de la calidad educativa en los resultados?

La respuesta es que no nos va muy bien. Para entender la razón, consideremos el caso de la posición socioeconómica (o SES), que universalmente se reconoce como un predictor de los resultados obtenidos en las pruebas. Posición socioeconómica es un término algo amorfo acuñado por los sociólogos para describir la ubicación social de individuos y hogares. No es raro que se utilice alguna medida del ingreso como si fuera una medida de la posición socioeconómica, pero en realidad esta implica algo más que eso y se mide de manera apropiada por medio de una combinación del ingreso, el logro educativo de los padres y su estatus laboral. Las tres variables predicen adecuadamente los resultados, incluso cuando se toman por separado. Por ejemplo, en la NAEP realizada en 2000 para la materia de matemáticas de doceavo grado, el desempeño promedio de los estudiantes que reunían los requisitos para recibir almuerzo gratuito o a bajo costo se encontraba a 0.71 desviaciones estándar por debajo del resultado promedio de los otros alumnos. Esa es una diferencia considerable, casi tan grande como la que se observa en matemáticas de octavo

grado entre Estados Unidos y Japón y equivale a casi tres cuartas partes de la diferencia promedio típica entre los estudiantes afroamericanos y blancos.

Dado que los chicos de posición socioeconómica (SES) alta por lo general no asisten a las mismas escuelas que los niños de SES baja, para aislar los efectos de la calidad de la escuela se necesita separarlos de los efectos de la SES. Los esfuerzos para hacerlo son innumerables, pero se encuentran con dos obstáculos: la medición de la posición socioeconómica suele ser muy deficiente, pero incluso cuando se mide bien no nos dice lo que se necesita saber acerca de los factores sociales que influyen en los resultados.

En las bases de datos que incluyen resultados obtenidos en las pruebas, la SES a menudo se calcula a partir de las medidas que estén disponibles o que puedan obtenerse con facilidad, las cuales por lo general son insuficientes. Como expliqué en el capítulo 5, la información que pueden recabar los distritos escolares y los estados acerca de las características de los estudiantes es muy limitada, y la información reunida acerca de la posición socioeconómica es deficiente incluso en la mayoría de los programas de evaluación a gran escala. (La NAEP depende de los estudiantes para obtener esta información pero –otro hallazgo de la abuela– los chicos no proporcionan información precisa acerca de muchos aspectos de los antecedentes familiares.)[4] Por lo tanto, a menudo tenemos que conformarnos con sustitutos poco sólidos de los datos que en verdad queremos de la SES, por ejemplo, tenemos que usar indicadores como el hecho de que los estudiantes cumplan los requisitos para obtener almuerzos gratuitos o a bajo precio y sus cálculos del número de libros en sus hogares. A pesar de la debilidad de esos sustitutos, su predicción de los resultados suele ser buena, aunque no lo suficiente. Los científicos sociales siguen discutiendo acerca de qué tan adecuado puede ser el trabajo si este combina una cantidad suficiente de esas medidas sesgadas, pero eso no ayuda a la

mayoría de los usuarios de los resultados de las pruebas (políticos, periodistas, padres) que por lo general sólo tienen acceso a los datos insuficientes recogidos por los estados y los distritos locales.

Un problema más importante, a menudo ignorado por los científicos sociales, es que, de cualquier manera, no es la posición socioeconómica lo que en realidad debe medirse. Incluso si se mide a la perfección, la posición socioeconómica es en sí misma un sustituto de las variables que tienen una influencia directa en el logro del estudiante y que por lo general no son medidas. Por ejemplo, dentro de ciertos límites, el ingreso en sí no influye directamente en el aprendizaje de los alumnos, sino otros factores que son influidos por el ingreso o que lo determinan. Mayores ingresos permiten a las familias brindar a sus hijos mejor nutrición y cuidado de la salud, mayor acceso a recursos educativos dentro y fuera del hogar, como lecciones de música y sesiones de tutoría fuera de la escuela. Los factores que permiten a los padres obtener un mayor ingreso –como niveles educativos más altos, motivación de logro y aceptación de la gratificación demorada– también pueden ayudar a sus hijos a lograr más en la escuela. Y el logro educativo de los padres, que en muchos estudios se ha demostrado que es un predictor importante del desempeño académico de sus hijos, es en sí mismo un indicador de otras cosas. El grado académico de la madre no implica una ventaja automática para sus hijos en las pruebas de logro. Sin embargo, los padres con un nivel educativo elevado a menudo se comportan de manera diferente a los padres con menor educación en muchas maneras que propician el logro: hacen un uso mayor y más diverso del lenguaje, dan más valor al éxito académico, etcétera. Por supuesto, hay muchos padres con poca educación (por ejemplo, mis abuelos inmigrantes) que hacen justamente las cosas que se necesitan para fomentar el logro académico y muchos padres educados que no las hacen, pero la probabilidad de encontrar esos tipos de conducta se incrementa con la mayor educación de los padres.

En innumerables estudios se ha demostrado la complejidad de los factores no educativos que pueden influir en el logro y los resultados obtenidos en las pruebas. Por ejemplo, un alentador estudio realizado a inicios de la década de los noventa, durante la llegada a Estados Unidos de los refugiados del Sudeste Asiático, intentaba descubrir los factores que pudiesen explicar por qué los hijos de inmigrantes pobres de Indochina superaban a sus pares del vecindario por un margen muy grande. Algunos de los resultados más sorprendentes fueron hallazgos de la abuela: por ejemplo, los padres de las comunidades inmigrantes por lo general enfatizaban la importancia del logro educativo como camino a una vida mejor, esperaban que sus hijos estudiaran de manera sistemática, a pesar de su precario estilo de vida trataban de proporcionarles un lugar y horario regular para estudiar, etcétera.[5] Este estudio me recordó una anécdota escrita por mi madre hace algunos años. Durante la Gran Depresión, mi abuela, una inmigrante que nunca tuvo oportunidad de recibir educación superior, preparó comida para algunos trabajadores huelguistas en la ciudad y le pidió a mi madre que se las llevara. Mi madre se rehusó, no porque estuviera en desacuerdo sino por lo que veía como el potencial de una gran vergüenza: igual que muchos adolescentes, no quería destacar entre la multitud. Objetó que llevar comida a los huelguistas no iba a resolver el problema fundamental. Mi abuela respondió: "Esa es la razón por la que tú vas a estudiar, para que puedas cambiar las condiciones. Mientras tanto, lleva la sopa".

Un segundo estudio, tan depresivo como inspirador era el de los inmigrantes, describe con un detalle verdaderamente doloroso las enormes diferencias en los ambientes lingüísticos proporcionados a los niños en dos comunidades casi adyacentes: una empobrecida y con educación deficiente y la otra con gran educación y mucho más próspera. Los investigadores encontraron diferencias notables en la frecuencia y tipos de interacción verbal en ambas comunidades, las cuales fueron igualadas por enormes diferencias en las

tasas típicas de adquisición de vocabulario por parte de los niños.[6] El simple hecho de controlar el nivel de ingreso de las poblaciones escolares no captura plenamente esos tipos de influencias en los resultados que obtienen los estudiantes en las pruebas.

Un tercer principio para identificar las causas de los resultados de las pruebas es un caso especial del segundo: *la simple coincidencia en el tiempo no establece causalidad*.* En los años recientes, a pesar de las asombrosas negaciones de algunas dependencias, la evidencia ha mostrado con claridad que el bióxido de carbono se ha ido acumulando en la atmósfera, la temperatura ha ido en aumento y las capas de hielo polar se han derretido. Al mismo tiempo han aumentado también los resultados en matemáticas, pero lo más seguro es apostar a que no se debe dar el crédito de ello al derretimiento de las capas polares. Usted diría que nadie en su juicio atribuiría el aumento de los resultados en matemáticas al derretimiento de las capas polares, pero eso implicaría que no se entiende lo esencial. A menudo una coincidencia *parece* más plausible y esto lleva a la gente a elaborar argumentos causales que en realidad no tienen más base que esta ilustración absurda. Si usted sigue la cobertura que se hace en la prensa de la evaluación educativa, con frecuencia encontrará explicaciones de los resultados de las pruebas que se basan en una simple coincidencia, por ejemplo, afirmaciones de que una política es (o no es) exitosa porque los resultados de las pruebas se incrementaron (o no lo hicieron). Esas explicaciones pueden ser erróneas o ser correctas, pero siempre son injustificadas: sin información adicional no podemos saber si son acertadas.

---

* El segundo principio se refiere a una correlación "transversal", una relación que aparece en un solo punto en el tiempo. En un determinado momento, los niños que hacen más tarea también tienden a obtener mayores calificaciones. Este tercer principio se refiere también a una correlación, pero esta ocurre a lo largo del tiempo. Un ejemplo sería un incremento que se da a lo largo del tiempo en la cantidad de tarea realizada, lo cual se acompaña por un incremento en las calificaciones.

Por último, surge un cuarto principio cuando se intenta discernir los efectos de la calidad educativa en los resultados obtenidos en las pruebas: *no se debe confiar en los resultados obtenidos en pruebas de alto impacto si no se cuenta con otra evidencia que los confirme*. Es frecuente que los resultados obtenidos en las pruebas de alto impacto (las que tienen consecuencias de consideración para alumnos o maestros) sufran una inflación considerable. Es decir, los incrementos en los resultados obtenidos en esas pruebas suelen ser mucho mayores que las verdaderas mejoras en el aprendizaje de los alumnos. Peor aún, esta inflación es sumamente variable e impredecible, por lo que no hay manera de saber en qué escuela se inflaron los resultados y en cuál son legítimos. Hay varias razones para eso, incluyendo la preparación inapropiada para la prueba y la simple trampa. En el capítulo 10 se analiza este problema, sus causas y su gravedad.

El problema que esto genera para la evaluación de las escuelas es evidente: ¿cómo pueden identificarse las escuelas y los programas eficaces, los que merecen ser reconocidos y emulados, si no se puede saber cuáles son las escuelas cuyos altos resultados son falsos? Eso tampoco es nuevo. Algunos expertos en medición, de manera notable George Madaus de Boston College, lo anticiparon hace por lo menos cuatro décadas, y en un reporte publicado hace 20 años por la Oficina de Presupuesto del Congreso advertí que el problema estaba creciendo y que eso afectaría nuestra capacidad para supervisar las tendencias en el desempeño de los estudiantes; desde 1991 la advertencia ha sido confirmada por varios estudios empíricos cuidadosos y por lo menos desde 1988 ha sido discutido en los medios de comunicación públicos y en las publicaciones educativas especializadas, aunque con una frecuencia mucho menor de lo que se amerita. No obstante, la inflación de los resultados se ignora de manera rutinaria cuando se utilizan para identificar las escuelas o sistemas escolares eficaces, no sólo por los

políticos y la prensa sino también por los educadores y por algunos científicos sociales que deberían estar mejor informados.

Este cuarto principio es un poco diferente de los otros porque requiere que se cuestione la información descriptiva en lugar de ser prudente respecto a saltar de los datos de una prueba digna de confianza a una inferencia acerca de sus causas. También podría considerarse diferente porque parece ser específico a los resultados de las pruebas, pero en realidad no lo es; la inflación de las medidas se presenta también en muchas otras áreas, como las emisiones de los camiones de diésel y la contención del costo del cuidado de la salud. Una vez más, veremos más de esto en el capítulo 10.

Causas de algunos patrones importantes en los resultados de las pruebas

A menudo (como en casi todos los ejemplos anteriores) las personas buscan explicaciones de los resultados de una prueba en particular que, por la razón que sea, es importante para ellas. Los padres quieren identificar buenas escuelas para sus hijos; los políticos quieren reclamar el crédito por iniciativas exitosas o usar los bajos resultados para justificar las reformas; los periódicos quieren destacar supuestas diferencias en la calidad escolar; los críticos desean identificar los fracasos; los educadores quieren reclamar los éxitos, y así sucesivamente.

Pero el debate público acerca de la educación y otras políticas sociales es determinado por la preocupación por los patrones amplios en el desempeño de los estudiantes: tendencias seculares del país, ascendentes o descendentes, en los resultados de las pruebas; las grandes diferencias entre los grupos raciales/étnicos y económicos; otros aspectos de la variación en el desempeño de los estudiantes; diferencias en las puntuaciones promedio entre países, etcétera.

Esos patrones proporcionan un contexto para la discusión, ya que determinan la manera en que la gente entiende los problemas y sus expectativas de mejoras razonables. Las supuestas explicaciones de los patrones suelen desempeñar un papel crucial en esos debates. Aquí considero la evidencia relevante para dos de los patrones más importantes: las tendencias generales en los resultados de las pruebas en el curso de las cuatro décadas pasadas y la enorme variabilidad en el desempeño de los estudiantes estadounidenses.

Las tendencias en los niveles logrados en las pruebas a lo largo del tiempo pueden ser los resultados más analizados de los programas de evaluación a gran escala. Han sido un ingrediente básico del debate público y político por más de un cuarto de siglo, y gran parte de la discusión se enfoca en sus supuestas causas. La mayor parte de las explicaciones ofrecidas, sean o no correctas, no son sustentadas: es común que se ignoren los principios descritos y que se presenten los resultados como única evidencia de los cambios en la calidad de la educación.

Esto fue tal vez más sorprendente a principios de la década de los ochenta, cuando un intenso debate público se enfocó en la disminución generalizada de los resultados de las pruebas que ocurrió durante las décadas de los sesenta y los setenta, en el final del descenso y en las mejoras relativas mostradas por estudiantes de los grupos minoritarios. Los científicos sociales y otros comentaristas ofrecieron explicaciones sorprendentemente diversas para dichos patrones. Señalaron tanto factores sociales y educativos como otros factores no educativos, pero estos últimos tuvieron mucha menos fuerza en el debate público. Entre las explicaciones presentadas estaba la disminución del trabajo requerido en los cursos de la escuela secundaria; un tiempo insuficiente dedicado a las materias principales; la supuesta baja calidad académica y la mala preparación de los maestros; la creciente proporción de estudiantes que vivían en familias monoparentales; cambios en el tamaño de la

familia (la investigación sugiere que en muchos ambientes sociales, los niños de familias grandes obtienen menores puntuaciones); un incremento en el tiempo dedicado a ver televisión; la lluvia radiactiva de las pruebas atómicas (un estudio argumentaba que el deterioro era peor en los lugares en que la lluvia era más pesada); cambios en la composición étnica de la población estudiantil; la abolición de la segregación racial y programas de educación compensatoria como *Head Start* y *Title I* (todos ellos presentados como la causa de las mejoras relativas de las minorías); así como la evaluación de competencia mínima y las políticas conservadoras de la década de los ochenta (factores sugeridos como la causa del fin del deterioro). Y esta lista no es exhaustiva.

La mayoría de esas explicaciones, independientemente de que fueran correctas –o, para ser precisos, "parcialmente correctas" ya que ningún factor parece ser suficiente para explicar esas tendencias–, se presentaron sin mucha evidencia que las sustentara. A menudo, la única justificación de una explicación era una simple coincidencia: pensamos que X cambió más o menos en el momento en que las puntuaciones disminuyeron, de modo que X debe haber sido la causa de esa disminución. Y, por supuesto, esta coincidencia por sí misma no nos dice mucho; más o menos al mismo tiempo que disminuyeron los resultados de las pruebas ocurrieron muchos cambios en todo tipo de cosas, y casi todos fueron por completo irrelevantes. Muy pocos comentaristas actuaron como un pediatra competente lo haría (o, para el caso, como un mecánico competente), buscando evidencia que pudiese descartar explicaciones alternativas. Pocos mencionaron incluso las alternativas.

Peor aún, muchos de esos comentaristas ni siquiera entendieron bien la sincronización y, por lo tanto, señalaron coincidencias que en realidad no existieron. Como el debate se enfocó excesivamente en el nivel de la preparatoria, en particular en la prueba SAT, los críticos tendían a concentrarse en explicaciones que fuesen

concurrentes con cambios en los resultados de las pruebas de esos grados. Esto ignoraba el fin anterior del descenso en los grados inferiores, como se describió en el capítulo 5. Tampoco se tuvo en cuenta el hecho de que el desempeño educativo al final de la preparatoria representa los efectos acumulados de 12 años de estudios, por lo que algunas de las causas relevantes (como los factores que ayudaron a terminar ese declive) deben haber precedido, en ocasiones por muchos años, al cambio actual en los resultados obtenidos en la preparatoria.

Estudios más cuidadosos de las posibles causas de esas tendencias en los resultados sugieren dos generalizaciones: al parecer participaron muchos factores, cada uno con una contribución modesta, y las causas parecen ser tanto sociales como educativas. La evidencia indica con fuerza que esto fue así tanto al inicio como durante el fin del deterioro de los resultados de las pruebas en las décadas de los sesenta y los setenta. La evidencia es de dos tipos: patrones amplios en las tendencias e investigaciones que exploran en detalle posibles causas específicas.

Los patrones amplios mostrados por las tendencias en los resultados no son en sí mismos suficientes para establecer las causas, pero ofrecen ciertos indicios útiles y descartan algunas explicaciones posibles. Un patrón es la generalidad verdaderamente notable de la disminución en los resultados. La caída ocurrió en el sistema escolar altamente descentralizado de todo el país, afectó a estudiantes de diversos tipos y edades, se presentó en escuelas privadas y públicas y ocurrió incluso en Canadá. Esto contradice a quienes pretenden imputar toda la culpa a las políticas educativas y plantea la posibilidad de que también importantes fuerzas sociales estén involucradas.

Un segundo patrón es la uniformidad del deterioro entre áreas temáticas. Creo que este tema fue planteado primero por el sociólogo Christopher Jencks. La lógica es sencilla y, para mi forma de pensar, convincente: si sólo hubiesen fallado las políticas y prácticas

educativas, uno esperaría ver un mayor deterioro en materias impartidas principalmente en la escuela (matemáticas) que en las que reciben mayor influencia de la "instrucción indirecta" fuera de la escuela, como lectura y en especial, vocabulario. No fue eso lo observado: no hubo diferencia sistemática en la disminución entre áreas temáticas. Lo que muestra un marcado contraste con las mejoras de las dos décadas anteriores, sorprendentes en matemáticas de la escuela primaria, modestas en matemáticas de la secundaria e insignificantes en lectura.

El patrón final es el efecto de la cohorte descrito en el capítulo 5: la disminución terminó primero en los grados iniciales (a mediados de la primaria) y el nadir ascendió gradualmente a medida que esas cohortes avanzaban en la escuela. Jencks fue también el primero en señalarlo: este tipo de efectos de la cohorte son más congruentes con algunas explicaciones sociales que con la mayor parte de las explicaciones educativas.[7]

La evidencia disponible acerca de supuestas causas específicas de las tendencias en los resultados de las pruebas es insuficiente para evaluarlas a todas ellas, pero es adecuada para descartar algunas y para calcular el tamaño de los efectos que otras pueden haber tenido. La evidencia sugiere que diversos factores sociales y educativos pueden haber contribuido a las tendencias, pero que ningún factor puede explicar más que una parte modesta del total. Por ejemplo, según mi cálculo, los cambios en la composición demográfica de la población estudiantil pueden haber explicado 10 o 20 por ciento del deterioro y haber desalentado un tanto el aumento posterior en los resultados. Los cambios en la composición familiar pueden haber contribuido al deterioro y al repunte. Algunos cambios en la práctica educativa, como la relajación en los requisitos y contenido de los cursos, pueden haber desempeñado un papel, sobre todo en los grados superiores. Los cambios en las tareas para casa tal vez tuvieron un efecto pequeño.

Las mejoras relativas de los estudiantes afroamericanos también parecen reflejar varios factores. La investigación sugiere que factores sociales, como un incremento en el logro educativo de los padres, contribuyeron a las mejoras pero son insuficientes para explicarlas por completo. Las evaluaciones también sugieren la posible contribución, aunque modesta, de programas como *Title I* y *Head Start* así como del fin de la segregación racial. La investigación por lo general encuentra que sus efectos a largo plazo, de ser positivos, son bastante pequeños y que el impacto de los dos primeros en las mejoras relativas de los afroamericanos es disminuido por el hecho de que muchos de los estudiantes atendidos por esos programas son blancos.[8]

Tomada entonces en conjunto, la evidencia indica que las tendencias en el desempeño de los estudiantes reflejan una confluencia complicada de diversos factores, tanto educativos como sociales. En retrospectiva, quizá este es también un hallazgo de la abuela. El sistema educativo estadounidense es complejo, incluso fragmentado, y el desarrollo de las capacidades cognitivas de los niños es un proceso de enorme complicación que apenas empezamos a entender. ¿Por qué deberíamos esperar que un solo factor tuviese un impacto muy grande en el desarrollo en una amplia variedad de contextos de toda la nación y que mientras tanto se mantuviera sin cambio el resto de las múltiples influencias sobre el aprendizaje de los estudiantes? O, para plantear esta misma pregunta en términos más amplios, ¿por qué deberíamos esperar que un solo paquete de reformas educativas suscite tendencias ascendentes en los resultados de las pruebas que sean incluso más grandes que el declive (lo cual ha sido la meta de numerosas políticas recientes)?

El debate simplista acerca del gran deterioro podría servir como una historia con moraleja, pero hasta ahora no lo ha hecho. El debate acerca de las causas del desempeño de los estudiantes no se ha vuelto más complejo –quizá *realista* sea un término más

adecuado– desde entonces. Eso puede apreciarse en las declaraciones acerca del supuesto éxito o fracaso de la ley NCLB. El programa se convirtió en ley apenas en enero de 2002, y se requirió mucho tiempo para que los estados respondieran –por ejemplo, para que compraran y aplicaran las pruebas adicionales requeridas por la ley. Si el programa tiene éxito se necesitará tiempo para sentir sus efectos. No obstante, cuando se hicieron públicos los resultados de la NAEP de 2004 (realizada más o menos dos años después de que se firmó la mencionada ley), los partidarios de la ley señalaron un incremento en los resultados de matemáticas como argumento del éxito de la política, mientras que los críticos hicieron referencia a la falta de incremento en los resultados de lectura para demostrar que no había tal éxito. No se advirtió en absoluto que los resultados de matemáticas habían estado aumentando a toda velocidad durante años anteriores a la promulgación de la ley y, por lo general, no se mencionó la posibilidad de que otra cosa (en realidad, muchas otras cosas) pudieran estar sucediendo al mismo tiempo. Peor todavía, algunos comentaristas no confiaron en absoluto en la NAEP y aceptaron sin cuestionar los incrementos en los resultados obtenidos en las pruebas estatales que los educadores estaban determinados (en algunos casos desesperados) por lograr.

Una segunda cuestión que se debate explícitamente con menos frecuencia pero que sin embargo de manera implícita juega un papel importante en el desarrollo de la política educativa es la siguiente: ¿qué ocasiona la enorme variabilidad en las puntuaciones mostrada por los estudiantes estadounidenses? Plantear esta pregunta es esencial para poder establecer objetivos de mejoramiento. ¿Qué significa decir que "todos los estudiantes pueden alcanzar estándares elevados"? Durante años, los reformadores y los responsables de las políticas educativas han repetido esta afirmación en distintas formas, como si fuera un mantra. Usted lo encontrará así en la explicación que hizo el Departamento de Educación de la

ley NCLB y en las declaraciones de los departamentos de educación de varios estados.

Cuando esta aseveración se convirtió en un ingrediente básico de la política educativa hace alrededor de 15 años, empecé a preguntar su significado a la gente que la utilizaba. Indiqué que la afirmación de que "todos los niños pueden alcanzar estándares elevados" podía significar al menos tres cosas. Primero, podía significar que la variación en el desempeño de los estudiantes seguiría siendo enorme, pero que la distribución completa se dispararía hacia arriba, de forma que los niños en el extremo inferior de la distribución terminarían por tener un mejor desempeño que el de la mayoría de los niños en la actualidad. Segundo, podía significar que la parte inferior de la distribución (todos los niños cuyos resultados se despliegan por debajo del promedio) se apretujaría hacia el promedio, o incluso más arriba, pero que se mantendría la dispersión entre los niños con altos resultados. Tercero, podía significar que todos se volverían más parecidos, de modo que también los niños con altos resultados se apretujarían hacia el promedio.

La reacción inicial a mi pregunta por lo general era de desconcierto. Luego de pensarlo por un momento, contestaban que esperaban una combinación de los dos primeros resultados: todos mejoran, pero los niños con bajos resultados mejoran todavía más, elevando sus puntuaciones en relación con el resto. Por supuesto, nadie quería decir que los niños de alto rendimiento se volverían mediocres, de modo que dejaríamos que sus resultados permanecieran dispersos.

Sin embargo, la distribución del desempeño es muy amplia, es decir, una brecha considerable separa a los niños con resultados bajos y elevados. Por lo tanto, si el objetivo de desempeño es alto o incluso apenas regular en relación con la distribución actual, para empujar a la mayoría de los niños por encima de ese objetivo se necesitaría que toda la distribución presentara un importante movimiento al alza y una compresión (que los niños localizados en

el fondo ascendieran en relación con los niños con mayores puntuaciones). Antes toqué el tema de las expectativas poco realistas de una mejora general. Aquí me gustaría considerar el problema de la compresión de la distribución del desempeño.

Muchos de los objetivos de desempeño establecidos en los años recientes parecen basarse en la expectativa de que la distribución de los resultados de la prueba puede reducirse de manera espectacular. A esto lo llamo el mito de la desaparición de la varianza. Si uno mira los datos, es difícil ver cómo podría comprimirse tanto la distribución. Podría reducirse, sí, pero no de manera espectacular.

Creo que el razonamiento que subyace a esas expectativas es la idea de que las injusticias educativas son la raíz de gran parte de la variación del desempeño en Estados Unidos. Sólo así tendría sentido argumentar que las reformas educativas pueden lograr que la distribución del desempeño se reduzca considerablemente.

No puede negarse que en Estados Unidos existe una desigualdad desmesurada, y en mi opinión inaceptable, en las oportunidades educativas. Existen muchas investigaciones que lo documentan, pero también esos son hallazgos de la abuela: documentan con horrible y sistemático detalle lo que no puede dejar de verse con sólo entrar en un puñado de escuelas. Y aunque existe un debate académico acerca de cuánto afectan el desempeño las desigualdades en recursos y oportunidades, el movimiento se demuestra andando o, como lo dirían los economistas, en las preferencias reveladas. La mayoría de las personas que recibieron una buena educación y conocen las escuelas no pondrían a sus hijos en la mayoría de los planteles que atienden sobre todo a niños con carencias.

A partir de este hecho inevitable, hay un salto al parecer pequeño a la suposición de que si reducimos esas desigualdades la variabilidad en el desempeño de los alumnos se achicará y el desempeño de algunos niños que se encuentran en el fondo se elevará hacia el promedio (o más allá). Y esta es de verdad una expectativa razonable.

Lo que no es razonable es esperar que a partir de entonces la variabilidad en el desempeño disminuya de manera notable. La razón es sencilla; hay muchas cosas además de las desigualdades educativas (como las flagrantes inequidades sociales) que también contribuyen a la variabilidad en el desempeño de los alumnos.

En el capítulo 5 ya se había mencionado una pieza de evidencia que apoya expectativas más modestas: los hallazgos de estudios internacionales como el TIMSS. Esos estudios no muestran una relación fuerte y sistemática entre la heterogeneidad educativa y social de los países y la variabilidad en los resultados de sus estudiantes. Por ejemplo, Japón y Corea tienen sistemas educativos más equitativos que Estados Unidos al nivel del octavo grado, pero el rango de resultados de sus estudiante es tan amplio como el de los alumnos estadounidenses (sólo que a un nivel mucho más alto).* Existe alguna variación entre países a este respecto, lo cual sugiere que debería ser posible reducir un tanto la variabilidad en el desempeño de los estudiantes. Pero el hecho de que la mayor parte de los países muestran variabilidad que no es muy diferente a la que se observa en Estados Unidos indica que deben establecerse expectativas modestas a este respecto.

---

* En la prueba TIMSS, pero no en la prueba PISA (que examina a estudiantes mayores y los agrupa por edad en lugar de hacerlo por grado), aparece un patrón interesante que es congruente con la opinión de que las oportunidades educativas son más equitativas en Japón y Corea. Si bien la variabilidad total de los estudiantes de esos dos países es similar a la que se observa en Estados Unidos, no sucede lo mismo con el desglose de esa variación. La variación *entre* las aulas de Japón y Corea es mucho menor que en Estados Unidos. Es decir, los resultados promedio de las aulas son más parecidos entre sí en Japón y Corea. Sin embargo, la variabilidad de sus estudiantes *dentro* del salón de clases promedio es en consecuencia más grande. Esto es lo que cabría esperar si las oportunidades educativas y los antecedentes sociales fuesen más similares entre las aulas de esos dos países que en las de Estados Unidos. El estado de Iowa también muestra menos variación entre las escuelas de lo que es común en Estados Unidos como un todo.

Otra pieza de evidencia proviene del examen del efecto de las diferencias entre grupos raciales y étnicos. Algunos de los síntomas mejor conocidos y más penosos de las desigualdades sociales y educativas en Estados Unidos son las grandes diferencias promedio en los resultados obtenidos por estos grupos, en particular entre los afroamericanos, los latinos y los blancos no hispanos (véase el capítulo 5). Hace algunos años ofrecí una conferencia en la que presenté el argumento, basado en la prueba TIMSS y en otros datos, de que no podíamos esperar una reducción importante en la variabilidad de los resultados de los estudiantes. Una disminución sí, pero no considerable. Algunas personas de la audiencia reaccionaron con una hostilidad que me dejó desconcertado. Después, uno de los que me interpelaron explicó el tenor de la reacción: muchos habían pensado que yo sostenía que las diferencias en el logro de grupos raciales y étnicos son fijas, tal vez innatas y que no pueden reducirse. Quedé conmocionado y horrorizado. Ese no era mi argumento en absoluto.

Desde entonces, cuando sé que este tema va a irrumpir en medio del debate, llego preparado con dos diapositivas que también utilizo cada año en uno de mis cursos introductorios. Empiezo planteando la siguiente pregunta: si pudiésemos borrar las diferencias promedio actuales en los resultados de las pruebas de los principales grupos raciales y étnicos en Estados Unidos, de forma tal que la distribución de los resultados de cualquier grupo fuese parecida a la distribución entre los blancos no hispanos, ¿cuánto se reduciría la variabilidad en los resultados –la variabilidad en la población total de estudiantes? Por lo regular nadie se atreve a hacer una conjetura; debe ser evidente que yo no haría la pregunta a menos que esperara que las suposiciones aventuradas fuesen erróneas. La respuesta: muy poco. Lo calculé cuatro veces: para lectura y matemáticas de octavo grado y usando los datos de la Evaluación Nacional del Progreso Educativo y del Estudio Longitudinal de Educación Nacional de 1988, uno de los poco frecuentes

estudios longitudinales del Departamento de Educación que son nacionalmente representativos. La reducción en la variabilidad (técnicamente la desviación estándar) oscila de alrededor de 0.5 por ciento a 9 por ciento en los cuatro casos.

¿Cómo puede ser ese el caso? Dado lo grande de las diferencias promedio entre esos grupos, es predecible a partir de la teoría estadística, pero es contraintuitivo. La razón es que la variabilidad *dentro* de cada grupo es muy grande y simplemente se traga el impacto de las diferencias promedio entre los grupos. Eliminar (no sólo reducir) las diferencias raciales y étnicas del grupo reduciría un poco la variabilidad total del desempeño de los estudiantes, y eliminar otras desigualdades, como la disparidad promedio en el desempeño entre blancos pobres y ricos, la reduciría un poco más. Pero al final, todavía enfrentaríamos una formidable variabilidad en el desempeño y todavía necesitaríamos un sistema educativo capaz de atender a los estudiantes de ese amplio rango de desempeño.

En muchos sentidos, entonces, un examen cuidadoso del panorama general da lugar a algunas de las mismas lecciones que un esfuerzo concienzudo por explicar los resultados obtenidos en una escuela o sistema escolar particular. El desarrollo del logro educativo es un proceso complejo en el que influye una gran variedad de factores, algunos de los cuales escapan al control de los educadores. Resulta poco realista la simple atribución de las diferencias en los resultados de las pruebas a la calidad de la escuela o, de igual manera, la simple suposición de que los resultados son en sí mismos suficientes para revelar la eficacia educativa. De manera más general, las explicaciones simples de las diferencias en el desempeño por lo general son ingenuas. Todo esto ha sido establecido por la ciencia. Iré un paso más adelante: esto sugiere que para ser realista suele ser necesario establecer objetivos más modestos para las mejoras en el desempeño de lo que muchos políticos recientes han estado acostumbrados a hacer. Al final del libro regresaré a este punto. ■

Capítulo 7
Error y confiabilidad: ¿qué tanto no sabemos de lo que estamos hablando?

Poco después de terminar el posgrado, trabajé en la evaluación de un programa federal diseñado para controlar los costos de los programas Medicare y Medicaid. En ese entonces yo era analista en la Oficina de Presupuesto del Congreso y el presidente de un subcomité del grupo legislativo, que albergaba dudas acerca del programa, había solicitado que se evaluara. El Departamento de Salud y Servicios Humanos, que administraba el programa y evidentemente quería conservarlo, preparó una evaluación muy similar a la nuestra. Se pidió a ambas instancias que presentaran los resultados de sus estudios en una sesión del subcomité.

Siendo Washington como es, estaba claro que no sería yo quien presentaría el testimonio, ya que tenía muy poca experiencia y mi rango era muy inferior al de la persona que haría la presentación por parte del Departamento de Salud y Servicios Humanos. Sin embargo, como era el analista del equipo con mayor conocimiento, escribí el testimonio y, como se acostumbra, acompañé a la persona que testificaría en la sesión en previsión de que se le hiciera una pregunta que no pudiese responder sin mi ayuda.

Las evaluaciones de la Oficina de Presupuesto del Congreso y del Departamento de Salud y Servicios Humanos fueron similares en sus resultados así como en su enfoque. El Departamento de Salud y Servicios Humanos concluyó que el programa estaba ahorrando una pequeña cantidad, algo así como 10 centavos por dólar. En contraste, nuestro reporte llegó a la conclusión de que el programa estaba perdiendo alrededor de 10 centavos por cada dólar.

El único problema era que ninguna de las evaluaciones podía apoyar una conclusión tan precisa como "por cada dólar gastado se ahorraron $1.10 (o $.90)". Los estudios se fundamentaron en diversas suposiciones y decisiones discutibles acerca de los métodos, además de que utilizaron datos que no eran idóneos y que sólo se basaban en muestras. Por lo tanto, cuando escribí nuestro testimonio no afirmé que el programa estuviera perdiendo dinero, sino que "dentro de cualquier estimación razonable del margen de error", las dos evaluaciones habían llegado más o menos a la misma conclusión: el programa quedaba más o menos a mano, ni ahorraba ni perdía una cantidad considerable de dinero. Sostuve que quienes desearan conservarlo o terminarlo tendría que buscar motivos distintos al costo.

La lacónica frase "dentro de cualquier estimación razonable del margen de error" atrajo más atención de lo que yo esperaba o deseaba. El presidente del subcomité se inclinó de su asiento en el estrado, miró fijamente a la persona que presentaba mi testimonio (mi jefe) y dijo algo parecido a "¿Qué demonios es esa cosa del 'margen de error'? ¿No significa eso que no tienen idea de lo que están hablando?".

Mi jefe ni se inmutó, por suerte, ya que yo era demasiado inexperto y ese arrebato me puso demasiado nervioso para atinar a responder. Sólo encontré una respuesta después de que había terminado la sesión y me encontraba a salvo fuera de la habitación "Sí, el 'margen de error' es una forma de cuantificar la medida en que no sabemos de qué demonios estamos hablando".

Este grado de incertidumbre, la segunda naturaleza no sólo de los científicos sociales sino de los muchos otros campos, es lo que se denomina *error* en la medición educativa y en el resto de las disciplinas estadísticas.

En la evaluación educativa el error adopta dos formas análogas pero distintas que se conocen como *error de medición* y *error del muestreo*.

La cantidad de error determina a su vez la *confiabilidad* de la calificación obtenida en una prueba: entre mayor sea el error menor es la confiabilidad. El reconocimiento de la importancia del error y de la confiabilidad ahora es tan general que en ocasiones se hace referencia a esos factores en la ley. Por ejemplo, la ley «Que ningún niño se quede atrás» de 2001, que en la actualidad es el precepto legal de mayor influencia en la evaluación de la educación básica y media-superior, establece que todas las evaluaciones que exige "deberán... utilizarse para propósitos para los cuales dichas evaluaciones sean válidas y confiables", y afirma que el desempeño de grupos de estudiantes en una escuela (como los estudiantes de grupos minoritarios) no debe considerarse por separado para calcular si el progreso anual de la escuela es el apropiado "en el caso de que el número de estudiantes en una categoría sea insuficiente para arrojar información que sea estadísticamente confiable".[1]

¿Qué significa en realidad "información confiable"? En el habla común, la palabra *confiable* puede significar muchas cosas, como el hecho de ser fidedigno, fiable o constante. No obstante, en la medición educativa el término tiene un significado específico de fundamental importancia. La clave para entender la confiabilidad es comprender lo que los expertos en medición y otras ciencias estadísticas entienden por "error".

Error de medición

Hace varios años, los padres cuyos hijos presentaron el examen del Sistema de Evaluación Integral de Massachusetts (*Massachusetts Comprehensive Assessment System*, MCAS) recibieron un reporte parecido al de la figura 7.1. Los resultados del estudiante eran indicados por una pequeña barra negra vertical (que según la escala ubicada en la parte inferior de la figura parece localizarse en una

Figura 7.1. Ejemplo de un informe del Departamento de Educación de Massachusetts para los padres de familia, MCAS *Tests of Spring 2002, Parent/Guardian Report*

¿Cómo le fue a _____ en esta prueba?

ÁREA TEMÁTICA	NIVEL DE DESEMPEÑO	RESULTADO ESCALADO	PRESENTACIÓN DEL RESULTADO Y RANGO DE RESULTADOS PROBABLES			
			Advertencia	Necesita mejorar	Competente	Avanzado
Matemáticas	Competente	280				

200 220 240 260 280

puntuación de alrededor de 247) y una barra horizontal más larga (que en el ejemplo se extiende más o menos de 240 a cerca de 255) a la que se asignó la leyenda "rango de resultados probables". Para facilitar la comprensión de esos informes, se entregó a los padres una guía en la cual se explicaba que "el desempeño del estudiante se muestra como un rango de resultados. La barra en esta sección muestra dónde cae el resultado de su hijo dentro de un rango de nivel de desempeño. La línea vertical en el centro de la barra representa el resultado escalado de su hijo... en cada prueba. La barra horizontal presenta el rango de resultados que su hijo podría recibir si presentara la prueba muchas veces".[2]

Esta barra horizontal, denominada en la figura "rango de resultados probables", representa el error de medición. Para ponerlo en los términos empleados hace años por el presidente del subcomité, el ancho de la barra cuantifica lo que no sabemos acerca del desempeño de este alumno particular en el Sistema de Evaluación Integral de Massachusetts a partir de este caso único de la evaluación. Todos los resultados obtenidos en la prueba tienen un rango de incertidumbre como este –algunos mayores, otros menores– y muchos otros programas de evaluación reportan esta información, aunque no siempre de manera similar a esta o con tantos detalles. Los esfuerzos de Massachusetts y algunos otros estados y distritos

por comunicar esta incertidumbre a los padres, la prensa y a otros son encomiables. Sin embargo, sé por experiencia que este tipo de información resulta desconcertante para mucha gente que no entiende ni las fuentes de la incertidumbre ni el significado del rango descrito.

En el lenguaje de la medición, ¿qué es en concreto lo que se entiende por "error"? En el habla cotidiana suele hablarse de error para referirse a una medición que de manera sistemática es equivocada. Por ejemplo, al hablar de una báscula del baño que por lo general arroja lecturas con alrededor de dos kilos de más, la gente podría decir: "En realidad he perdido más peso de lo que esta báscula indica ya que tiene un error de dos kilos". Las personas podrían usar la misma frase si se hubieran parado sobre la báscula una sola vez. Sin embargo, en ese caso no sabrían si la discrepancia de dos kilos se mantiene a lo largo del tiempo o fue una casualidad. Podría ser, por ejemplo, que la primera vez que se pararon sobre la báscula hubiera registrado dos kilos de más y que si se hubieran subido de nuevo la segunda lectura estuviera más cerca del blanco o muy por debajo.

Esta distinción aparentemente sutil entre la imprecisión constante y la inconstante reviste una importancia fundamental en la evaluación y, para el caso, en cualquier aplicación de la estadística. Los errores en esos campos (el error de muestreo y el error de medición) por lo general significan incertidumbre o imprecisión. Es decir, se refieren a una *inconsistencia*, no a una imprecisión sistemática. Esta incertidumbre es inherente a cualquier medición única, como el "rango de resultados probables" expuesto arriba. También se presenta en cualquier encuesta de opinión pública, como la que se revisó en el capítulo 2: otras encuestas habrían proporcionado estimaciones diferentes a la hecha por Zogby. La realización reiterada de las mediciones, como la aplicación repetida de las encuestas a diferentes muestras o la evaluación de un estudiante en

múltiples ocasiones, revelaría esta anomalía. En contraste, un error sistemático –como una báscula que de manera constante arroja una lectura dos kilos más alta– se denomina *sesgo*. En la evaluación educativa es difícil exagerar las implicaciones de esta diferencia.

Para concretar esta distinción, regresemos a las básculas de baño baratas. Si su báscula es de baja calidad, es probable que arroje contradicciones notables de una medición a la siguiente. Se sube a la báscula y lee que pesa 82.5 kilos; se baja y vuelve a subirse y esta vez pesa 82 kilos; lo intenta de nuevo y lee 82.75 kilos. La anomalía de la escala es, para todo propósito práctico, aleatoria e impredecible. Nunca se sabe si la siguiente lectura será más alta o más baja y no se sabe si la siguiente discrepancia entre dos lecturas será mayor o menor a la previa.

Esta inconsistencia es el error de medición. Una forma eficaz, aunque neurótica, de cuantificar el problema (averiguar en qué medida no sabe de qué diablos está hablando cuando se sube a la báscula una sola vez para calcular su peso) consistiría en subirse y bajarse de la báscula muchas veces, digamos, unas 100 veces. La distribución o dispersión de las 100 mediciones le daría una idea del grado de incertidumbre de cualquier medición única. De hecho, el "rango de resultados probables" de la figura 7.1 es una estimación similar: refleja la distribución de resultados que el estudiante obtendría de presentar muchas veces la prueba de matemáticas del MCAS.

Al no iniciado suele resultarle difícil de entender la diferencia entre el error de medición y el sesgo, pero un experimento sencillo puede ayudar a distinguirlos. Si realiza mediciones repetidas y su promedio se aproxima gradualmente a la respuesta correcta, se trata de un error de medición. Si el promedio de las medidas repetidas se mantiene incorrecto, incluso después de una gran cantidad de mediciones, lo que tiene es sesgo. Por ejemplo, dada la imperfección de la báscula barata, si sólo se pesó *una vez*, es probable que la única lectura sea demasiado alta o demasiado baja, y esta discrepancia de su

verdadero peso podría ser efecto del sesgo o del error de medición; si sólo se subió una vez a la báscula no puede saber de cuál se trata. La causa se aclararía si se subiera en varias ocasiones a la báscula y llevara el promedio de sus mediciones repetidas. Si la báscula tiene un error de medición, pero no está sesgada, el promedio se aproximará a su verdadero peso a medida que aumenta el número de lecturas. Las fluctuaciones aleatorias en las lecturas se cancelarían mutuamente. Sin embargo, si la báscula tiene sesgo, este promedio sería erróneo sin importar el número de veces que se suba a la báscula. En cualquier caso, a medida que aumenta el número de estimaciones, el promedio se estabilizará alrededor de un solo número, pero en el caso del sesgo el número seguiría siendo erróneo.

En la evaluación educativa el error de medición es mucho más importante que en el caso de la báscula de baño, y durante décadas sus causas y efectos han sido objeto de muchas investigaciones. Para la interpretación y empleo de los resultados resultan de particular importancia tres preguntas acerca del error de medición. ¿Qué es lo que causa esta inconsistencia? ¿Cómo se mide, es decir, cuando sólo tenemos una medición cómo cuantificamos lo que no sabemos acerca de lo que hablamos? Y, por cuestiones prácticas, ¿qué tanto importa?

En el capítulo 2 mencioné tres fuentes de falta de coherencia en los resultados obtenidos en las pruebas. La fuente de error que, por mucho, ha recibido más atención de las técnicas psicométricas es la variación en el desempeño de una muestra de reactivos a otra, ilustrado por *indolente* y *parsimonia* en el ejemplo de la prueba de vocabulario. Los reportes técnicos que acompañan a la mayor parte de las evaluaciones a gran escala por lo general incluyen estimaciones de confiabilidad (denominadas *consistencia interna* de la confiabilidad de los estadísticos) que sólo toman en consideración el error de medición que surge de la selección de los reactivos para elaborar esa forma particular del examen.

La segunda causa del error de medición se apunta en la explicación del rango de resultados probables ofrecido por el Departamento de Educación de Massachusetts: las fluctuaciones en el desempeño de los estudiantes a lo largo del tiempo. Si se aplicara la misma prueba a los alumnos en múltiples ocasiones (suponiendo que no recuerdan sus contenidos y no se cansan o se hartan del examen), obtendrán un resultado diferente de una ocasión a la siguiente por distintas razones como las variaciones en sus propias condiciones (enfermedad, sueño, ansiedad ante las pruebas y cosas por el estilo) y en los factores externos (condiciones del aula en que se realiza el examen, etcétera). Cuando los estudiantes presentan en muchas ocasiones la prueba SAT y encuentran una fluctuación en sus resultados, esa incongruencia refleja las dos fuentes de error: diferencias en los reactivos seleccionados para las formas sucesivas de la prueba y los días buenos y malos de los estudiantes.

Una báscula de baño también puede ilustrar esas dos fuentes del error de medición. El problema de obtener una estimación confiable de su peso no es sólo cuestión de anomalías aleatorias en la conducta de la báscula a las 6.30 a.m. o en una mañana determinada. También hay otro problema: su verdadero peso fluctúa de manera considerable de un día a otro, incluso si no hay una tendencia subyacente en el peso, como resultado de lo que ha comido y bebido, el ejercicio que ha realizado, etcétera. Esta es la razón del consejo común, aunque engañoso, de que si está tratando de perder peso no debería pesarse todos los días sino más bien de manera poco frecuente. Este es un mal consejo porque si sólo compara dos mediciones, digamos con una semana de distancia, la aleatoriedad de la conducta de su báscula y las fluctuaciones de su peso sumarán el error a ambas estimaciones, lo que creará un riesgo considerable de que la comparación sea por completo equivocada a menos que el cambio en su peso haya sido lo bastante grande para anular esas incongruencias. Un método mejor, aunque compulsivo, sería tomar mediciones frecuentes pero

ignorar las diferencias de una ocasión a la siguiente, en lugar de obtener promedios o buscar tendencias subyacentes.

La tercera fuente común del error de medición es la inconsistencia en la calificación de los resultados. Por supuesto, esto no aparece en el caso de las pruebas de opción múltiple calificadas por una máquina, a menos que algo funcione mal, como sucedió en un caso reciente, muy difundido, en que la humedad ocasionó que las hojas de respuesta de la prueba SAT se dilataran tanto que las burbujas no se alinearon adecuadamente con los sensores del escáner. Sin embargo, la falta de coherencia en la calificación puede ser importante en cualquier programa de evaluación que requiere que el trabajo de los alumnos sea calificado por personas, sobre todo cuando los productos que tienen que calificar son complejos. El problema aparece de manera rutinaria en la revisión de evaluaciones como las pruebas de redacción, las pruebas de ensayo y las evaluaciones de portafolios. Suele conocerse como acuerdo entre jueces, consistencia entre jueces o confiabilidad entre jueces. Como veremos a continuación, el último término es potencialmente engañoso.

Las variaciones en el proceso de calificación proporcionan un buen ejemplo de la distinción entre sesgo y error de medición. Dichas variaciones adoptan diversas formas. Una son las diferencias sistemáticas en la severidad de los jueces. Cuando estudiaba el posgrado, fui asistente de enseñanza de un grupo grande dividido en tres secciones, cada una de las cuales tenía como profesor a un estudiante de posgrado distinto. Poco después de iniciar el semestre, varios alumnos de mi sección se quejaron de que mi criterio para calificar los ensayos era sistemáticamente mucho más severo que el de los otros dos asistentes. Lo verifiqué y tenían razón (luego ajusté toda la distribución de mis calificaciones para que coincidieran mejor con las de las otras secciones). También es posible que con el paso del tiempo los calificadores cambien su

conducta, muchas veces de manera impredecible. Por ejemplo, un calificador, cuando tiene que calificar muchos ensayos, puede ponerse de mal humor y, por ende, más severo, mientras que otro quizá se vuelva más indulgente porque lo único que quiere es terminar el trabajo. Otro puede producir un rango de resultados progresivamente más estrecho a medida que pasa el tiempo. Incluso los calificadores cuya severidad es, en promedio, comparable pueden diferir en su benevolencia de un estudiante al siguiente. Por ejemplo, dos calificadores pueden ser influidos en distinto grado por los errores gramaticales o incluso por el descuido al escribir, como resultado de lo cual pueden clasificar a estudiantes específicos de manera diferente aunque su exigencia promedio sea la misma. También pueden dejarse llevar por características del estudiante distintas al trabajo que tienen a la mano, como su comportamiento en clase. Las rúbricas de calificación claras y la capacitación cuidadosa pueden disminuir, pero no eliminar, esas incongruencias. (Por ese motivo, todos los trabajos de mis grupos se califican de manera anónima, los estudiantes son identificados únicamente por su número de identificación y presentan sus exámenes en computadora para evitar problemas generados por la escritura. Sólo añadimos los nombres al final del semestre de modo que sea posible tomar en consideración otros factores al asignar las calificaciones finales, como la participación en clase y circunstancias atenuantes.)

Es claro que algunas contradicciones en el proceso de calificación no son aleatorias sino más bien sistemáticas. ¿Cuándo producen sesgo y cuándo generan sólo error de medición? Consideremos un ejemplo de cada uno. Primero, suponga que una estudiante presenta una prueba de admisión que debe calificarse a mano. Presenta la prueba una sola vez y, sin que ella lo sepa, el examen es calificado por A, un empleado de la empresa de evaluación. Entre los empleados de esa empresa, A es relativamente indulgente, de modo que la calificación de la estudiante es un poco más alta de lo

que habría sido si la hubieran calificado muchos de los otros empleados. La asignación del examen de la estudiante al calificador A fue, para todo propósito, aleatoria, de modo que si volviera a presentar el examen de nuevo es muy probable que no le tocara el mismo calificador la segunda vez. Por lo tanto, si presenta la prueba de manera repetida y luego se promedian sus resultados, las incongruencias en la calificación se compensarían y el promedio tendería hacia su "verdadero" resultado o, por lo menos, a su resultado depurado de los efectos de la falta de consistencia del juez. En consecuencia, podemos considerar que esta incongruencia es sólo error de medición que contribuye a una banda de incertidumbre que se extiende en ambos sentidos de su resultado verdadero, como en el ejemplo del MCAS en la figura 7.1.

Veamos otro caso: un programa de evaluación en que los exámenes de los estudiantes son calificados por sus maestros. Esta es la manera en que se califican los Exámenes Regentes del Estado de Nueva York, también fue el sistema utilizado en la década de los noventa para calificar los portafolios de escritura en Kentucky y se empleó por un tiempo breve cuando Vermont estableció las evaluaciones de portafolios de escritura y de matemáticas. Si el resultado obtenido en la prueba es lo bastante importante, los maestros tienen un incentivo para calificar de manera indulgente. ¿Qué sucede si sucumben a esta tentación? La distorsión en el resultado de una estudiante no se compensará si presenta el examen de manera repetida porque en cada ocasión sería calificada por la misma persona, su maestro. Esto es justamente lo que sucedió en el caso de la evaluación de portafolios de Kentucky. Cuando otros jueces volvieron a calificar muestras de los portafolios en una auditoría estatal de resultados, se descubrió que los resultados asignados por muchos maestros a sus propios alumnos eran demasiado altos.[3] Como esta distorsión no se eliminaba con la repetición de los exámenes, era sesgo y no error de medición. Por razones que nadie ha determinado,

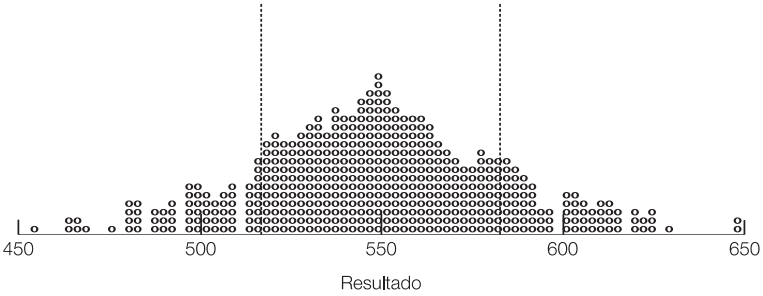
este tipo de sesgo no apareció en el programa de evaluación de portafolios de Vermont (hasta donde sé, no existen estudios sistemáticos recientes que indiquen si el Examen Regente del Estado de Nueva York tiene un problema de sesgo de la calificación).

Los usuarios de las calificaciones disponen de tres métodos diferentes para cuantificar o exponer el error de medición y la confiabilidad. El primero es ejemplificado por la barra horizontal de la figura 7.1. La barra expresa el error de medición como un rango de incertidumbre (240 a 255) en la misma escala que el resultado estimado del estudiante (247). Técnicamente, esto se cuantifica con un estadístico llamado *error estándar de medición (EEM)*.

Para concretar este enfoque, veamos otro ejemplo hipotético. Simulé unos datos para representar lo que sucedería si una estudiante, cuyo "verdadero" resultado fuese 550, presentara de manera repetida una prueba hipotética con la misma cantidad aproximada de error de medición que el contenido en las secciones verbal y matemática de la prueba SAT, la cual es una prueba de alta confiabilidad (es decir, en comparación con otras pruebas, la prueba SAT tiene relativamente poco error de medición y un error estándar comparativamente pequeño). La figura 7.2 muestra cómo luciría la distribución de resultados si la estudiante presentara la prueba 500 veces. (Esto supone lo imposible: que no hay efecto alguno de la práctica o la fatiga y que el tiempo para el examen y la paciencia son ilimitados.) Cada círculo representa un solo resultado.

Cualquier resultado individual que la estudiante hubiese obtenido si presentara la prueba una sola vez estaría muy lejos del resultado deseado de 550, y aunque esta es una prueba relativamente confiable, algunas de los resultados en la figura están demasiado alejados. Sin embargo, el resultado promedio de esas pruebas simuladas fue casi exactamente de 550, que es el valor "verdadero" para esta sustentante hipotética, el resultado purgado del error de medición. Como en el caso de la báscula de baño, el error aparente en

Figura 7.2. Error de medición en una prueba hipotética



los resultados individuales se eliminó al repetir las mediciones. (Si usted realizara este ejercicio de manera repetida, encontraría que no siempre se elimina del todo, pero el promedio casi siempre sería muy cercano a 550 debido al gran número de observaciones.)

Como muestra la gráfica, hay muchos resultados ligeramente por encima o por debajo de 550, mientras que las puntuaciones que se alejan mucho de ese nivel son relativamente escasas. Esto se traduce en la probabilidad de que nuestra pobre sustentante, si sólo se le permitiera un intento, obtendría una puntuación muy alejada de su verdadero resultado de 550. Tiene una probabilidad bastante alta de obtener cualquier calificación en un rango de, digamos, 25 o 35 puntos de 550, aunque es poco probable un resultado alejado más de 50 puntos.

Esta figura ilustra la respuesta a la pregunta planteada hace décadas por el furioso presidente del subcomité: demuestra cómo cuantificamos qué tanto no sabemos acerca de lo que estamos hablando (en este caso, de un solo resultado obtenido en una prueba). Las líneas punteadas representan una distancia de un error estándar de medición por arriba o por debajo del promedio. En este caso, el rango es de 66 puntos (33 en cada dirección de la media) que es similar al error estándar de medición de la prueba SAT. Alrededor de dos tercios de las observaciones simuladas caen dentro de ese rango. Esto es así en general: un examinado con cualquier resultado

verdadero dado, que presenta una prueba una vez, tiene una probabilidad de alrededor de dos tercios de obtener una puntuación dentro del rango de un error estándar de medición por debajo y uno por arriba, y una probabilidad de uno en tres de obtener un resultado alejado en más de un error de medición estándar de su verdadero resultado. Si se extiende el rango a dos errores estándar por arriba y por debajo de la media (en este caso, de 484 a 616) se incluiría alrededor de 95 por ciento de los resultados que podría obtener nuestra infatigable examinada. El rango de incertidumbre reportado puede ser en más o menos ya sea con uno o dos errores estándar (no sé cuál fue el caso en el ejemplo de la figura 7.1). Este es un "margen de error", análogo (pero como veremos, no idéntico) al que hizo saltar hace algunos años al presidente del subcomité.

Un aspecto de este error que resulta inquietante para algunas personas es que incluso con una banda amplia (digamos, más o menos dos errores estándar), no existe certeza de que el resultado verdadero en realidad se encuentre dentro de la banda.* Sólo podemos decir que probablemente se encuentra dentro de esa banda. No se trata de un problema exclusivo de la medición educativa, se presenta en todas las inferencias estadísticas. Por ejemplo, es raro que un solo estudio pueda decir con absoluta certeza que un nuevo medicamento es eficaz en el tratamiento de una enfermedad específica. Más bien, lo común es que la investigación indique que una diferencia en los resultados entre un grupo tratado con el nuevo medicamento y uno que no recibió el tratamiento es tan grande que es muy poco proba-

* Esta forma de presentar el error es técnicamente incorrecta, pero como la descripción adecuada es engorrosa lo usual es que la gente utilice esta descripción más sencilla. En estricto sentido, lo que estamos calculando no es la probabilidad de que el resultado verdadero que deseamos se encuentre dentro de algún rango del resultado estimado obtenido por el estudiante. Más bien, estamos dando respuesta a la siguiente pregunta: si el resultado que se observó inicialmente en realidad es el resultado verdadero, ¿cuál es el rango de resultados que obtendríamos si se repitiera la prueba?

ble que ocurriese por azar y que, por ende, es un resultado probable del tratamiento. Con el tiempo, podemos acumular tanta evidencia de tantos estudios que la incertidumbre se vuelve insignificante, pero en la práctica, siempre acecha en el fondo. En el caso de una sola prueba educativa, la incertidumbre rara vez es insignificante.

Lo anterior ilustra una de las maneras comunes de cuantificar lo que no sabemos acerca de un solo resultado pero, para la mayoría de los lectores, esto no responderá la tercera pregunta planteada antes: ¿Este rango de incertidumbre es lo bastante grande para preocuparnos? Para responder a esta pregunta necesitamos saber qué tanto se dispersan los resultados en la escala particular utilizada. Por ejemplo, a menos que sepamos algo sobre la variabilidad de los resultados en la escala en que se reportan los resultados del MCAS, no hay forma de saber si "el rango de resultados probables" de 15 puntos que se muestra en la figura 7.1 es grande o pequeño. En algunas pruebas (como la ACT, que tiene una puntuación máxima de 36 en cada prueba), 15 puntos serían un rango de incertidumbre enorme. En otras (como la prueba SAT que tiene un rango de 200 a 800 en cada escala), 15 puntos serían un margen de error muy pequeño.* Es posible que el ejemplo hipotético de la figura

* Técnicamente, el problema no es el rango total de resultados (ya que ese rango puede extenderse de manera arbitraria) sino su variabilidad, que por lo general se mide con la desviación estándar. La desviación estándar (descrita en el capítulo 5) es una medida de variabilidad tal que si los resultados se ajustan a la curva de campana, alrededor de dos tercios de todas las puntuaciones caerán en el rango de una desviación estándar por debajo a una desviación estándar por arriba de la media. (La desviación estándar y el error estándar de la medición son análogos: la primera cuantifica la variabilidad en un conjunto de resultados –por ejemplo, los resultados de un grupo de muchos estudiantes– mientras que el último cuantifica la variabilidad de un conjunto de medidas múltiples de una persona). La desviación estándar de los resultados en la parte de matemáticas de la prueba SAT es de 115 puntos, mientras que en la parte de matemáticas de la prueba ACT es de 5 puntos. Por lo tanto, un rango de 15 puntos es pequeño en relación con la variabilidad que exhiben los estudiantes en la prueba SAT, pero es enorme (23 veces más grande) en relación con su variabilidad en la prueba ACT.

7.2 tenga más significado para los lectores familiarizados con la prueba SAT porque se elaboró de modo que fuese muy parecido a esa prueba en términos de la escala y la confiabilidad. Para esos lectores resultará evidente que el margen de error en la prueba SAT no es insignificante –no es trivial tener una posibilidad de incluso 5 por ciento de obtener un resultado de más de 66 puntos por arriba o por debajo del resultado verdadero. Y repito: la prueba SAT está muy bien construida y es altamente confiable. La incertidumbre es mayor en el caso de muchas otras pruebas educativas.

El impacto práctico de este rango de incertidumbre depende de cómo se utilicen los resultados. Supongamos que la prueba hipotética de la figura 7.2 es un examen de admisión a la universidad. Si se fuera a establecer un punto de corte fijo (digamos que en este ejemplo el corte es de 545) y se rechazara a cualquiera que obtuviese una puntuación menor a 545, incluso un margen de error muy pequeño tendría consecuencias muy serias para los estudiantes (como la de nuestro ejemplo) cuyo resultado verdadero se encuentra cerca del punto de corte. Esos estudiantes tendrían una probabilidad bastante alta de ser incorrectamente rechazados o aceptados por el simple hecho de que el error de medición ocasionó que sus resultados fueran algo más bajos o más altos de lo que deberían haber sido. Sin embargo, si los resultados obtenidos en la prueba se utilizaran sólo como una pieza de información que contribuye a la decisión de admitir o rechazar a los estudiantes, una cantidad modesta de error de medición tendría poco impacto. Esta es una de varias razones por las que el College Board recomienda a los funcionarios de los departamentos de admisión que tomen en cuenta la confiabilidad y el error de medición cuando utilicen los resultados de la prueba SAT y que los traten como "indicadores aproximados" de las fortalezas del estudiante. En particular les sugiere que empleen los resultados junto con otras fuentes de información sobre las capacidades de

los estudiantes (como las notas o calificaciones en la escuela y las exposiciones escritas), que no tomen decisiones basadas en pequeñas diferencias en los resultados y que no impongan una punto de corte mínimo a menos que se use junto con otra información.[4]

La confiabilidad suele presentarse mediante el uso de un segundo estadístico, el *coeficiente de confiabilidad*, que es útil para los expertos pero difícil de entender para el público en general. El coeficiente de confiabilidad, a diferencia del error estándar, no se expresa en la escala de la prueba. Independientemente de la prueba o de la escala en que se coloquen los resultados, el coeficiente de confiabilidad siempre varía de 0 (un resultado que no es otra cosa que error de medición, es decir, ruido aleatorio) a 1 (un resultado perfectamente sistemático, sin error de medición alguno). Esto hace que el coeficiente de confiabilidad sea comparable de una prueba a la siguiente, incluso cuando los resultados estén expresados en escalas distintas. Este coeficiente tiene también otras propiedades matemáticas que lo hacen útil para quienes realizan evaluaciones técnicas de los resultados obtenidos en las pruebas, y es el estadístico de confiabilidad que más a menudo se reporta en la práctica. Sin embargo, a pesar de todas sus ventajas técnicas, el coeficiente de confiabilidad tiene un inconveniente importante: a diferencia del error estándar de medición, no informa de manera directa a los usuarios no entrenados la cantidad de error inherente al resultado.

Los usuarios de los resultados reciben a menudo reglas prácticas para decidir qué tan altos deberían ser los coeficientes de confiabilidad, pero son arbitrarias y puede ser más útil contar con estándares de comparación. En los programas de evaluación a gran escala, las pruebas más confiables tienen coeficientes de confiabilidad en el rango de .90 o un poco más altos. Por ejemplo, las confiabilidades de consistencia interna de las secciones matemática y verbal de la prueba SAT (las estimaciones de confiabilidad que sólo toman en cuenta el error de medición que surge del muestreo

de los reactivos) se encuentran en el rango de .90 a .93, dependiendo de la forma.[5] Los exámenes de lectura y de matemáticas de las Pruebas de Habilidades Básicas de Iowa, uno de los instrumentos comerciales de logro más antiguos y más utilizados, contienen incluso menos error de medición por la selección de reactivos y coeficientes de confiabilidad de consistencia interna un poco más elevados.[6] Algunas pruebas estatales desarrolladas por encargo tienen niveles de confiabilidad igualmente altos. En 2003, se calculó que la confiabilidad de la consistencia interna de la prueba de matemáticas para décimo grado del MCAS, la cual debe aprobarse para obtener el diploma de preparatoria, era de .92.[7] Por lo general, la confiabilidad es menor en el caso de pruebas más cortas y subpruebas (por ejemplo, la puntuación en la parte de cálculo de las pruebas de matemáticas de nivel primaria), lo mismo que en muchas evaluaciones que son innovadoras o inusuales. Por ejemplo, en 2001, una aplicación anterior de la evaluación alterna del estado de Washington para los estudiantes con graves discapacidades encontró confiabilidades de consistencia interna en el rango de .72 a .86 además de un considerable error de calificación.[8]

Si bien esas comparaciones ayudan a mostrar cuál es el nivel de confiabilidad que se debería esperar, no explican qué tan buenos o malos son los números. La respuesta es que incluso cuando el coeficiente de confiabilidad es alto, persiste un error de medición importante. Existen varias maneras de aclarar esto. Una es preguntar, para un determinado coeficiente de confiabilidad, ¿qué tan grande es la banda de error alrededor de un resultado individual? Por ejemplo, aun cuando la prueba SAT es muy confiable, con un coeficiente de confiabilidad mayor a .90, el error estándar de medición es de más de 30 puntos, similar al mostrado en la figura 7.2. Un coeficiente de confiabilidad de .80 indica una banda de error alrededor de 40 por ciento más grande y un coeficiente de confiabilidad de .70 indica un error estándar de medición casi 75 por

ciento mayor al de la figura 7.2. Otra forma consiste en preguntar, dado un determinado coeficiente de confiabilidad, ¿qué tan bien puede predecirse un segundo resultado si se conoce el primero? Si se tiene el primer conjunto de resultados de un grupo de estudiantes, un coeficiente de confiabilidad de .90 indica que esos primeros resultados permitirán predecir alrededor de 80 por ciento de la variabilidad de los segundos resultados. Con un coeficiente de confiabilidad de .70, sólo puede predecirse más o menos la mitad de la variabilidad del segundo grupo de resultados.

Una tercera manera de mostrar la confiabilidad, que es particularmente pertinente para la evaluación basada en estándares que en la actualidad domina las evaluaciones exigidas por el estado, es la consistencia de la clasificación o decisión basada en los resultados. Esto sólo surge cuando los resultados obtenidos en una prueba se descomponen en categorías discretas, como con un solo punto de corte de aprobación-reprobación o el pequeño número de categorías creado al informar los resultados en términos de unos cuantos estándares de desempeño, como el criterio de "competente" exigido por la ley NCLB. Este último tipo de reporte ahora es general y ya era común incluso antes de promulgar la mencionada ley. Cuando el desempeño se informa de este modo, uno puede preguntar: si se clasifica a un estudiante en una categoría (digamos, no competente o competente) con base en los resultados de una prueba ¿qué tan probable sería que se le reclasificara en la otra categoría si se le examinara en una segunda ocasión? Entre más error de medición haya en la prueba –es decir, entre menos confiable sea el resultado–, mayor es la probabilidad de que la clasificación se modifique de una ocasión a la siguiente.

La inquietante respuesta es que esa clasificación varía más a menudo de lo que a uno le gustaría. La tabla 7.1, que se adaptó de un estudio realizado por dos científicos sociales en Rand para la Universidad de la Ciudad de Nueva York, muestra el porcentaje de

Tabla 7.1. Porcentaje de estudiantes cuya categoría de aprobado/reprobado cambiaría en una segunda evaluación en varias combinaciones de tasa de aprobación y confiabilidad

Porcentaje de aprobación	Confiabilidad		
	.70	.80	.90
90	11	9	6
70	22	17	12
50	26	21	14
30	22	18	13
10	11	09	6

Fuente: Adaptado de Stephen P. Klein y Maria Orlando, *CUNY's Testing Program: Characteristics, Results, and Implications for Policy and Research* (Santa Monica, CA: Rand, 27 de abril de 1999), Tabla 2.

alumnos que recibirían una clasificación diferente de aprobación o reprobación en dos instancias del examen. Las hileras indican diferentes niveles de los estándares, con los más severos hacia el fondo. La hilera superior representa un estándar muy indulgente que permitiría aprobar a 90 por ciento de los estudiantes; la segunda hilera, un estándar que permitiría aprobar a 70 por ciento de los alumnos, y así sucesivamente. Las tres columnas representan diferentes niveles de confiabilidad, medidos por el coeficiente de confiabilidad. Sigamos con la columna de .90, que se acerca al mejor escenario. La tabla muestra que a menos que se establezca un estándar muy bajo o muy elevado, un número importante de estudiantes sería reclasificado si se repite la prueba. Si se establece el estándar cerca de la mitad, de modo que apruebe de 30 a 70 por ciento de los alumnos, en una segunda aplicación del examen se clasificaría de manera diferente de 12 a 14 por ciento de los estudiantes.

¿Qué tan importante es esta inconsistencia? Una vez más depende de cómo se utilicen los resultados, pero para algunos usos importaría mucho. Si el resultado único se utilizara por sí solo

para tomar una decisión importante –como negar el diploma de preparatoria o rechazar la solicitud de ingreso a la universidad de un estudiante–, incluso la inconsistencia inherente a una prueba de gran confiabilidad sería preocupante. Por ejemplo, en el caso de una tasa de aprobación de 50 por ciento en una prueba con una confiabilidad de .90, la mitad de los estudiantes clasificados de manera no consistente (7 por ciento de todos los que presentaron la prueba) reprobarían la primera vez pero aprobarían si se repite el examen. Esta es una razón por la que muchos estados que imponen como requisito para graduarse de preparatoria la obtención de una calificación permiten que los estudiantes repitan la prueba, a menudo en varias ocasiones.

Los estadísticos de confiabilidad que se proporcionan con los resultados de las pruebas tienden a subestimar el problema del error de medición porque a menudo sólo toman en consideración una o dos de las fuentes de error. Por consiguiente, el error es subestimado (porque se ignora el que proviene de algunas fuentes) y en la misma medida se sobreestima la confiabilidad. Por ejemplo, es común proporcionar a los usuarios de los resultados de las pruebas estimaciones de consistencia interna de la confiabilidad que toman en cuenta el error del muestreo de los reactivos pero que no reflejan la inconsistencia a lo largo del tiempo o, cuando es pertinente, incongruencias en el proceso de calificación.* La mayoría de los ejemplos presentados antes son de este tipo.

En el caso de las evaluaciones que son difíciles de calificar, en ocasiones se encontrarán reportes en que los estadísticos que

* Las estimaciones de la consistencia interna de la confiabilidad pueden identificarse a menudo por el nombre aunque no se describan como tal. La estimación que se reporta con más frecuencia es el *Alfa de Cronbach*. Los coeficientes de confiabilidad de *Kuder Richardson* son en esencia lo mismo, pero se aplican sólo a los reactivos de opción múltiple o a otros reactivos que se califican simplemente como bien o mal.

representan la consistencia del proceso de calificación (a menudo con la etiqueta engañosa de "confiabilidad entre jueces") se presentan sin otra información acerca del error y se tratan como si representaran la confiabilidad de los resultados, aunque ignoran las fluctuaciones a lo largo del tiempo y los efectos del muestreo de reactivos. Hace algunos años, en respuesta a esta última distorsión, H. D. Hoover, en ese entonces autor principal de la Prueba de Habilidades Básicas de Iowa, una batería de pruebas de logro de opción múltiple, comentaba que si la consistencia entre jueces es suficiente para indicar la confiabilidad de los resultados obtenidos en una prueba, entonces los resultados de la Prueba Iowa son perfectamente confiables porque no tienen error del proceso de calificación: las máquinas de escaneo óptico arrojarían calificaciones idénticas todas las veces que se introdujeran las hojas de respuesta.

¿Qué hace al resultado de una prueba más o menos confiable? Cuando es requerido calificar, mejorar la congruencia del proceso de calificación reducirá el error total de medición y aumentará la confiabilidad. Esto puede lograrse con un cuidadoso diseño y evaluación de las rúbricas que usan los jueces para calificar el trabajo de los estudiantes, una capacitación rigurosa de los jueces y la supervisión del proceso de calificación para detectar y corregir los problemas (por ejemplo, realizando una segunda calificación de una muestra aleatoria de trabajos). Los procedimientos para estandarizar la aplicación de la prueba ayudarán a disminuir las fluctuaciones en el desempeño de una ocasión a la siguiente.

Otro factor que influye en la confiabilidad de los resultados es la consistencia del contenido de la prueba, conocida como su consistencia interna (de ahí el nombre de los estadísticos de confiabilidad analizados antes) u *homogeneidad*. Recuerde que en la prueba de vocabulario la elección de palabras (indolente frente a parsimonioso) ocasionó variaciones en el ranking de los estudiantes que presentaron la prueba. Considere ahora una prueba de

matemáticas de cuarto grado. Se trata de un dominio amplio y la selección de reactivos causará algunas fluctuaciones en el desempeño, y por ende algún error de medición. Pero suponga que diseñó la prueba para que sólo incluyera problemas de resta de dos dígitos sin llevar, presentados en el formato vertical, por ejemplo:

$$\begin{array}{r} 57 \\ -25 \\ \hline 32 \end{array}$$

Como todos los posibles reactivos de este tipo son muy parecidos, haría muy poca diferencia cuáles eligiera y la fluctuación resultante en los resultados (el error de medición) sería pequeña. Sin embargo, el costo sería considerable: habría diseñado una prueba extremadamente limitada que podría ser útil como una prueba semanal sorpresa en una clase de matemáticas pero sería sumamente engañosa como una prueba de matemáticas de cuarto grado. Este es otro de los inevitables compromisos en la medición. Al diseñar una prueba de un dominio grande como el de matemáticas de cuarto grado, uno querría una cobertura razonablemente amplia del dominio para apoyar las conclusiones en que se está interesado, pero la amplitud del contenido reducirá la confiabilidad. Como sugieren los coeficientes de confiabilidad presentados antes, con un trabajo cuidadoso, los autores de las pruebas de dominios amplios pueden obtener altos niveles de confiabilidad de consistencia interna, aunque esto se ve restringido por la amplitud de la prueba.

Una de las influencias más importantes en la confiabilidad es la extensión de la prueba. Al analizar el ejemplo de la báscula de baño barata, señalé que si se pesaba las veces suficientes y promediaba las lecturas, el error de medición de las lecturas individuales se eliminaría y el promedio será una medida bastante buena de su verdadero peso. En la evaluación educativa los reactivos individuales de una prueba son análogos a las lecturas individuales de su

báscula. Entre más reactivos incluya en la prueba, más se compensará en cada uno de ellos el error de medición y más confiable será el resultado que se obtenga de la prueba. Esto implica también un compromiso: las pruebas más extensas son más costosas y, lo que es más importante, requieren más tiempo para su aplicación y, en consecuencia, implican mayores trastornos y reducción del tiempo que se podría dedicar a la instrucción.

Considerar tanto la extensión como la homogeneidad de la prueba ayuda a esclarecer una de las discusiones recientes más polémicas acerca de la evaluación educativa: la mejor mezcla de formatos que debe utilizarse. Desde finales de la década de los ochenta, los educadores, reformadores y expertos en medición han mostrado un interés generalizado por ir más allá de los formatos de opción múltiple que habían dominado la evaluación del logro desde el final de la Segunda Guerra Mundial hasta ese momento. Como se indicó en el capítulo 4, los formatos adicionales trataban de ser diversos e incluían reactivos que requerían respuestas escritas cortas; reactivos que exigían ensayos más largos; tareas de desempeño práctico (como tareas de ciencia que implicaban la realización de un experimento u otras manipulaciones de aparatos); evaluaciones de portafolios basadas en el trabajo generado en el curso de la instrucción, así como tareas híbridas que demandaban esfuerzo de grupo e individual. En general, si no se incrementa el tiempo para la prueba, el uso de formatos más complejos a menudo (aunque no de manera invariable) disminuye la confiabilidad. Existen dos razones relacionadas. Primero: entre más complejas sean las tareas, es probable que sean menos homogéneas. Por ejemplo, es probable que dos tareas prácticas en ciencia difieran en muchos aspectos que no son por fuerza fundamentales para los atributos del desempeño sobre el cual se sacarán las conclusiones. Segundo: dado que las tareas complejas se llevan más tiempo, habrá menos tareas por hora del tiempo del

examen y, por ende, menos oportunidad de eliminar el error de medición. Esto también es un compromiso porque, aun cuando disminuyen la confiabilidad, los formatos complejos pueden ofrecer otras ventajas importantes, como la posibilidad de evaluar habilidades que no es sencillo medir con formatos como el de opción múltiple.

Error de muestreo

La gente está menos familiarizada con el error de medición que con una segunda y muy similar forma de error: el error de muestreo. Mientras el error de medición es la incongruencia que surge de una medición de una persona a la siguiente, el error de muestreo es la inconsistencia que surge de la selección de personas (escuelas o distritos) particulares de los cuales se hace una medición. Cuando los científicos o los periódicos escriben sobre el error, por lo general se refieren a un error de muestreo, y fue la preocupación por el error de muestreo que dio lugar a mi comentario sobre el margen de error lo que llevó hace años al presidente del subcomité a decirle a mi jefe "¿No significa eso que no saben de qué diablos están hablando?".

Uno de los ejemplos más comunes del error de muestreo son los resultados de sondeos y otras encuestas. Cualquiera que lea un periódico los meses previos a una elección se encuentra con los resultados de frecuentes sondeos, escritos por lo general de una forma como esta: "En el sondeo más reciente de [el nombre de una organización encuestadora], 52 por ciento de los probables votantes dijeron que votarían por el candidato X si la elección fuese hoy. Treinta y ocho por ciento dijo que votaría por el candidato Y y 10 por ciento dijo estar indeciso. El sondeo encuestó a 658 probables votantes y tuvo un margen de error de más o menos 3 puntos por-

centuales". Incluso el presidente del subcomité que décadas atrás se enfureció por mi testimonio habría entendido este; después de todo, los políticos viven y mueren por los resultados de los sondeos.

El margen de error reportado por la mayoría de las ecuestas refleja inconsistencias que surgen del muestreo de las personas a las que se entrevista. Cada vez que se saca una muestra para entrevistar, habrá fluctuaciones aleatorias en los tipos de personas que responden. Un día, es posible que se pille a unos cuantos conservadores adicionales, el día siguiente a algunos progresistas de más, y así sucesivamente. La consecuencia será alguna fluctuación de los resultados. La incertidumbre se presenta incluso si se toma la muestra una sola vez (y de hecho los estadísticos pueden calcular el error de muestreo de una sola muestra si conocen el diseño del muestreo) pero la extracción de muestras repetidas lo hace evidente.

Por supuesto, los sondeos y otras encuestas también padecen sesgos que no se eliminan con las muestras repetidas. Los resultados de una encuesta pueden sesgarse si las preguntas están mal planteadas, aunque sea por accidente. Puede presentarse sesgo si la muestra de participantes no está bien diseñada para que sea representativa, por ejemplo, si el diseño da por resultado que se incluyen en la muestra demasiados votantes ancianos o con educación superior. Pero otro sesgo más insidioso puede surgir de la falta de respuesta. No todas las personas con las que hace contacto un encuestador aceptan responder. (¿Cómo respondería usted a las llamadas telefónicas de los encuestadores durante la cena?) La gente que se niega a participar es sistemáticamente distinta a quienes aceptan responder, y esas diferencias pueden crear un sesgo considerable en los resultados a menos que la tasa de respuesta (la proporción de individuos contactados que acepta participar) sea muy alta. En la práctica, es raro que los medios no especializados presenten esas tasas de respuesta. Cuando lo hacen, en ocasiones verá que son muy bajas, lo cual debería ser una advertencia para tomar

los resultados con reserva o incluso para ignorarlos del todo. Pero esto nos lleva a otra cosa: el problema que yo quiero aclarar aquí es el error de muestreo y su pertinencia para la evaluación educativa.

Por lo general el error de muestreo no era una preocupación importante en la evaluación educativa. Hasta hace muy poco tiempo el propósito principal de la mayoría de los programas de evaluación era calcular la competencia de los individuos, de ahí que al evaluar el error la atención se centraba sobre todo en el error de medición –la confiabilidad de la estimación de los resultados de los individuos. Hace décadas, cuando mis padres recibieron mis resultados de la Prueba de Habilidades Básicas de Iowa, el error de muestreo no era un problema (la calificación se refería sólo a mí), pero sí lo era el error de medición. En los años recientes esto ha cambiado por la rápida evolución de los usos de los resultados de las pruebas. Como expliqué en el capítulo 4, en la actualidad un uso importante de los resultados de las pruebas es describir y evaluar a grupos: escuelas, distritos y estados enteros.

Esto da lugar al error de muestreo: la inestabilidad en esos resultados agregados que resulta del muestreo de los estudiantes. La segunda cita de la ley NCLB presentada al inicio de este capítulo es un reconocimiento explícito del error de muestreo en la estimación de los cambios anuales en el desempeño.

El error de muestreo en los resultados agregados de las pruebas, como en todos los estadísticos basados en la muestra, es una función del tamaño de la muestra. Entre más gente se interrogue en una encuesta acerca de la probabilidad de voto, menor será el margen de error, o por lo menos el error que surge únicamente del muestreo aleatorio de los participantes.* Esto es parecido al impacto de la extensión de la prueba en el error de medición. Una

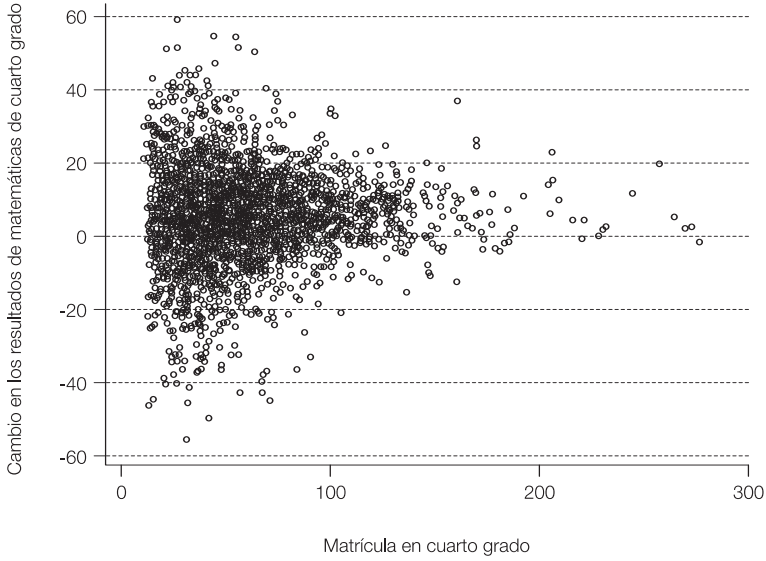
* En particular para una media, el margen de error disminuye como función de la raíz cuadrada del número de observaciones.

prueba más larga contiene más reactivos y, por ende, más oportunidad de eliminar las diferencias aleatorias entre reactivos (error de medición). Una muestra más grande incluye más individuos y, por ende, ofrece más oportunidad para eliminar las diferencias aleatorias entre las personas (error de muestreo).

El error de muestreo puede ser muy grande, lo cual es ilustrado en los estudios realizados por mi colega Tom Kane en colaboración con Douglas Staiger de la Universidad de Dartmouth. En el periodo que precedió a la promulgación de la ley NCLB, Kane y Staiger publicaron una serie de trabajos en que investigaron la confiabilidad de las medidas agregadas basadas en los resultados de las pruebas. Encontraron que los simples resultados agregados son muy poco confiables en muchas escuelas porque el número de estudiantes examinados es pequeño. En la figura 7.3 se muestra un ejemplo, el cual se adaptó de uno de sus estudios.[9] Esta gráfica expone el cambio anual en el resultado promedio obtenido en la prueba de matemáticas de cuarto grado aplicada en California entre 1999 y 2000. El eje vertical de la gráfica representa el cambio en el resultado promedio. Cada círculo representa el cambio en el resultado promedio de una sola escuela. El eje horizontal de la gráfica muestra la matrícula de cuarto grado en la escuela.

En la figura se aprecia que los resultados de las escuelas más grandes (es decir, las que tienen más de 200 alumnos en cuarto grado) son bastante estables y muestran muy poco cambio en cualquier dirección de 1999 al 2000. Sin embargo, son muy pocas las escuelas primarias con tantos estudiantes en un grado. La mayoría, como lo manifiesta la masa de círculos hacia el lado izquierdo de la gráfica, tienen matrículas relativamente pequeñas, a menudo con sólo uno, dos o tres grupos en ese grado. Para esas escuelas, en particular las que sólo tienen uno o dos grupos, los cambios anuales son muy variables y muchos de los planteles exhiben un cambio considerable en una u otra dirección. Otros estudios

Figura 7.3. Cambio en un año en los resultados promedio de las pruebas obtenidos por alumnos de cuarto grado en matemáticas en la evaluación STAR de California. Adaptado con autorización de los autores, tomado de Thomas J. Kane y Douglas O. Staiger, "The Promise and Pitfalls of Using Imprecise School Accountability Measures", *Journal of Economic Perspectives*, 16. Núm. 4 (2002): 91-114.



confirman que esos cambios son erráticos de un año al siguiente. Por ende, en su mayor parte esos cambios no representan modificaciones significativas en el desempeño de las escuelas pequeñas sino un error de muestreo más grande cuando se examina a relativamente pocos alumnos. En otras palabras, no es realmente el caso que el desempeño de muchas escuelas pequeñas mejore o se deteriore con rapidez mientras el desempeño de las escuelas más grandes se mantenga estable.

El problema del error de muestreo en los resultados agregados se complica por la exigencia de la ley NCLB de informar los resultados desagregados, es decir, los resultados de las pruebas de

grupos específicos de estudiantes, como minoritarios y discapacitados. Tal como aclararon algunos de los defensores de esta legislación, el propósito de este requisito era obligar a las escuelas a prestar atención al desempeño de los grupos que históricamente obtenían bajos resultados y hacer imposible que las escuelas salieran del problema mejorando sólo el desempeño de los alumnos con mejores puntuaciones, a quienes es más sencillo enseñar. Aunque he sido abiertamente crítico de muchos aspectos de la mencionada ley, tanto yo como algunos otros críticos del estatuto estamos de acuerdo en que esta presión es importante y existe por lo menos evidencia anecdótica que ha llevado a muchas escuelas a prestar más atención al desempeño de los grupos de alumnos con mayor riesgo de obtener bajas calificaciones.

El inconveniente es que las estadísticas reportadas para esos grupos, que suelen ser mucho menores que las poblaciones escolares mostradas en la figura 7.3, suelen ser muy poco confiables. Por consiguiente, las tendencias de corto plazo en esos resultados pueden ser engañosas y el aparente progreso o falta de progreso a menudo será ilusorio. Esta preocupación es lo que suscitó el lenguaje normativo al inicio del capítulo. La ley NCLB reconoce este problema y permite omitir el reporte de grupos que sean lo bastante pequeños para que su desempeño resulte poco confiable, pero deja en buena parte a los estados la decisión de cuántos estudiantes deben ser incluidos en un grupo (es decir, qué tan confiables deben ser los resultados) antes de exigir un informe separado para ese grupo.

Kane y Staiger demostraron otra consecuencia menos evidente de este requisito: entre más grupos tenga una escuela que deban reportarse por separado, más probable es que la escuela fracase debido únicamente al error de muestreo. La ley NCLB exige que cada grupo reportado debe hacer progresos anuales adecuados (AYP) para poder acreditar a la escuela el progreso anual. Entre

más grupos se reporten por separado, más probable es que al menos uno no logre hacer el progreso anual debido simplemente al azar, en particular el error de muestreo. Por ejemplo, digamos que en un año dado, una escuela tuvo un par de estudiantes muy exitosos de grupos minoritarios y un par de estudiantes con discapacidades muy poco exitosos. Supongamos también que, como suele suceder, esos patrones reflejan el error de muestreo: por azar, este año el subgrupo minoritario recibe resultados relativamente elevados y el grupo discapacitado resultados relativamente bajos en comparación con otras cohortes de esos grupos en la misma escuela en otros años. El cambio anual en el desempeño de los estudiantes del grupo minoritario parecerá algo mejor de lo que sería sin esos pocos estudiantes, y el cambio para los estudiantes con discapacidades parecerá algo peor.

Al informar los resultados globales de toda la escuela, el error de muestreo en los resultados de esos dos grupos se compensará. Sin embargo, el reporte desagregado no permite que sus resultados se compensen. Como la ley NCLB exige que todos los grupos reportados logren el progreso anual, la escuela no podría cumplir este objetivo debido a esos alumnos específicos con discapacidades, incluso si su desempeño global es bueno y el fracaso de ese grupo es resultado del error de muestreo.

Algunos lectores pueden preguntarse cómo surge el error de muestreo si todos (o casi todos) los alumnos de una escuela son examinados. Después de todo, en el caso de las encuestas los errores de muestreo se presentan porque sólo se cuenta con las respuestas de un pequeño porcentaje de las personas que en realidad van a votar. Este no es el caso con la mayoría de los programas de evaluación, los cuales en condiciones ideales examinan a casi todos los alumnos de un grado.

Esta pregunta fue tema de cierto debate entre los miembros de la profesión hace apenas unos años, pero ahora existe un acuerdo

general de que el error de muestreo es un problema incluso si se examina a cada estudiante. La razón es la naturaleza de la inferencia que se basa en los resultados de las pruebas. Si la inferencia concerniente a cada escuela de la figura 7.3 fuese *acerca de los estudiantes particulares de esa escuela en ese momento*, el error de muestreo no sería un problema porque casi todos ellos fueron examinados. Es decir, el muestreo no sería una preocupación si la gente utilizara los resultados para llegar a conclusiones como "los alumnos de cuarto grado que estuvieron en esta escuela en el año 2000 obtuvieron mayores resultados que el grupo concreto de estudiantes que estuvieron inscritos en 1999". Sin embargo, en la práctica es raro que los usuarios de los resultados se preocupen por esto. Más bien se interesan en conclusiones sobre el desempeño de las *escuelas*. Para esas inferencias, cada cohorte sucesiva de estudiantes inscritos en la escuela son sólo otra pequeña muestra de los alumnos que tendrían la posibilidad de inscribirse, de la misma manera en que las personas entrevistadas para una encuesta constituyen una pequeña muestra de quienes podrían haber sido entrevistados.

Un investigador me contó una anécdota que ejemplifica muy bien la importancia del muestreo al interpretar los resultados de las pruebas. En una reunión realizada hace algunos años, los profesores de una pequeña escuela de Maryland estaban desconcertados por un descenso notable en los resultados que duró un solo año en cada grado y avanzó un grado cada año, muy parecido a cómo podría verse el movimiento de una rata recién consumida a lo largo de una serpiente pitón. Una profesora explicó: "Ese es Leo". Se refería a un alumno problemático que se las arreglaba para bajar el desempeño de cada grupo en que estaba. No sé si tenía o no razón en los detalles, pero su explicación era razonable al señalar al error de muestreo como el probable culpable. Cuando cuento esta historia en alguno de mis grupos que incluyen a muchos antiguos maestros, la mayoría ríe con complicidad y el "efecto Leo" se

ha convertido en una manera conveniente de referirnos a las fluctuaciones a corto plazo en los resultados agrupados ocasionadas por el error de muestreo.

Significancia estadística

Los informes de los resultados de las pruebas a menudo son acompañados por declaraciones acerca de la "significancia estadística" (o la falta de la misma) de los hallazgos, digamos, una mejora en los resultados en comparación con las del último año o la diferencia en los resultados de dos distritos o estados. Esas referencias también aparecen en el reporte de muchos otros hallazgos científicos, como los resultados de la investigación sobre los efectos de un medicamento. No queda claro cómo entienden muchos periodistas lo que esa frase significa o lo que esperan que hagan los lectores en respuesta. Un periodista que llevó uno de mis cursos durante una licencia de su trabajo me proporcionó como ejemplo uno de sus artículos acerca de las diferencias en los resultados obtenidos en una prueba. Había escrito el artículo antes de llevar el curso y me lo mostró para analizar cómo podría expresar las cosas de manera diferente en el futuro. Había incluido en el artículo una sola oración que advertía que algunas de las diferencias no eran estadísticamente significativas: Sin embargo, una vez que entendió el error de medición y el error de muestreo, decidió que la referencia a la significancia era una cuestión de forma y que la había escrito sin tener una expectativa clara de que los lectores supieran qué hacer con la información.

El término *significancia estadística* es sólo otra manera de cuantificar qué tanto desconocemos acerca de lo que estamos hablando. Vamos a subirnos de nuevo a la báscula de baño barata. Suponga que está intentando perder peso. Al inicio de su régimen de dieta

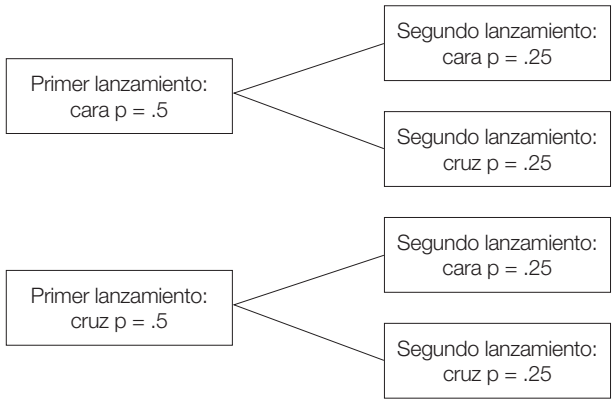
o ejercicio se pesa y obtiene una lectura de 65 kilos. Una semana después vuelve a pesarse y obtiene una lectura de 64 kilos. De acuerdo con su báscula ha perdido un kilo.

Pero ¿debería creerlo? Tal vez la aparente pérdida de peso es sólo el resultado del error de medición. Es decir, debido al error de medición, la aparente mejora sobre la báscula *puede ser el resultado del azar*. Si la báscula es en verdad mala y contiene mucho error de medición, la probabilidad de obtener por puro azar una "mejora" de un kilo es alta. Entre más confiable sea la báscula menos probable será que un determinado resultado aparezca sólo por azar.

Un ejemplo completamente distinto proviene de una clase que preparé en una ocasión para la materia de matemáticas en la secundaria de uno de mis hijos, una clase que no se me permitió impartir a pesar de la invitación de la maestra debido al intenso bochorno que podría haber causado. (Los padres de los estudiantes de secundaria entenderán que el bochorno no tenía nada que ver con la preocupación de lo bien que pudiera haberla enseñado.) Para este ejemplo usted tiene que imaginar que es un chico de 12 años algo renuente a retar abiertamente al profesor. Suponga que me paro enfrente del grupo y explico que voy a lanzar una moneda de manera repetida y a escribir los resultados en el pizarrón. Sólo yo puedo ver los resultados del lanzamiento de la moneda, por lo que usted tiene que aceptar mi palabra sobre ellos. La tarea es decidir después de cada lanzamiento si los resultados son tan improbables que usted estaría dispuesto a pasar al estrado y declarar abiertamente que supone que los estoy amañando, sea mediante el uso de una moneda cargada o simplemente mintiendo. Dejemos que p sea la probabilidad de la serie de resultados que obtengo. En el primer lanzamiento sale cara, lo cual tiene una probabilidad de .5 (uno de cada dos), por lo que todavía no hay motivo de hablar claro. En el segundo también sale cara, y la probabilidad de dos caras en serie es una de cuatro o .25. Puede

ver lo anterior en la figura 7.4: la mitad de las veces se obtendrá cara en el primer ensayo y, de esa mitad, la mitad llevará de nuevo a obtener cara en el segundo lanzamiento. La regla general de probabilidades independientes (si la moneda no está sesgada o cargada, la probabilidad de obtener cara en el segundo lanzamiento es independiente del resultado en el primero) es que la probabilidad de una serie de eventos igualmente probables es p^n, donde n es el número de eventos. De modo que con dos lanzamientos no hay todavía motivo para hacer cualquier cosa; la probabilidad de obtener este resultado con una moneda no cargada es bastante alta. El tercer lanzamiento también resulta ser cara, y la probabilidad de esta serie de caras es $.5^3$ o .125. Los resultados se vuelven más improbables pero todavía no ameritan el costo potencial de llamar farsante al maestro. El cuarto lanzamiento también es cara, $p = .5^4 = .0625$. Esto es más peliagudo. El quinto lanzamiento: también cara, $p = .5^5 = .0313$. Este es un resultado sumamente improbable. Si los lanzamientos fuesen en verdad al azar y yo repitiera este ejercicio una y otra vez, sólo obtendría esta secuencia de cinco caras en alrededor de tres veces de cada cien.

Figura 7.4. Las probabilidades de todas las combinaciones de dos lanzamientos de una moneda no sesgada


Primer lanzamiento: cara p = .5
Segundo lanzamiento: cara p = .25
Segundo lanzamiento: cruz p = .25
Primer lanzamiento: cruz p = .5
Segundo lanzamiento: cara p = .25
Segundo lanzamiento: cruz p = .25

En cierto punto, alguien del grupo diría que esos resultados son demasiado improbables para ser creíbles y que piensa que el lanzamiento de la moneda está amañado. En otras palabras, es *improbable que esta secuencia de resultados surgiera sólo por azar*, en este caso, la posibilidad aleatoria de una serie de lanzamientos. El estudiante *no sabe* que los resultados no surgieron por azar, pero son tan improbables que argumentará que es *probable* que el resultado se debiera a una cosa distinta al azar (en este caso, al hecho de que hice trampa).

La significancia estadística es sencillamente una declaración sobre la probabilidad de que cualquier resultado en cuestión pudiera haber ocurrido sólo por azar debido al error de muestreo, al error de medición o a ambos. Entre menor sea esta probabilidad, mayor confianza se puede tener en la explicación que supone alguna causa distinta al azar y mayor será la significancia estadística. Por convención, la mayoría de los científicos establecieron como umbral mínimo una probabilidad (la *p* en mi lanzamiento de la moneda) menor de 0.05, pero esto es una simple convención. (Al inicio esto puede resultar confuso: entre *menor* sea la probabilidad de que los resultados surgieran por azar, *mayor* es su significancia estadística y más confianza tenemos en que algo distinto al azar causara los resultados.) Las personas de mi grupo que argumentaran a favor de un efecto de la trampa habrían cruzado en el quinto lanzamiento de la moneda el umbral convencional para la significancia estadística.

¿Qué debería usted hacer entonces con esta información al encarar los resultados obtenidos en la prueba? Existen dos errores comunes que deben evitarse. El primero es que los resultados estadísticamente significativos son reales y que los resultados no significativos (por lo general, los estadísticos no los llaman "insignificantes") no lo son. Por desgracia, estamos jugando con la suerte, un hecho que puede ser bastante perturbador en algunos casos de inferencia estadística, como cuando se tiene que decidir si se

debe recetar un determinado tratamiento médico. Un hallazgo que es estadísticamente significativo tiene menor probabilidad que uno no significativo de haber surgido por azar. Sin embargo, todavía existe la posibilidad (aunque pequeña) de que un hallazgo estadísticamente significativo *se debiese* al azar. También es posible que un hallazgo estadísticamente no significativo *no surgiera* por azar. Entre menos información se tenga (entre menos confiable sea la prueba o más pequeña la muestra de personas) más probable es que seamos engañados por una falta de significancia estadística.

Para concretar lo anterior, la figura 7.5 presenta el resultado promedio de matemáticas de cuarto grado de varios estados en la Evaluación Nacional del Progreso Educativo realizada en el año 2000. Cada estado es representado por una muestra aproximada de 2 500 sustentantes. Esta figura es la esquina superior izquierda de una figura mucho más grande que expone los resultados para cada estado que participó en la evaluación. A esta y a otras figuras similares de la NAEP se les conoce a menudo como "gráficas de pantimedia" porque el área blanca en la parte central se parece a la corrida de una media. Cada hilera y columna representan a un estado. Por ejemplo, la hilera superior y la columna del extremo izquierdo representan a Minnesota (MN), que ese año fue el estado que obtuvo el mayor resultado. La segunda columna e hilera representan a Montana (MT), el estado con el segundo resultado más alto, etcétera. Al leer hacia abajo de la primera columna (Minnesota) verá que MN, MT y KS están en blanco, mientras que el resto de las celdas por debajo de esas tres están sombreadas. Cualquier comparación en este rango blanco no es estadísticamente significativa. Así, por ejemplo, aunque Minnesota superó a Kansas, la diferencia pudo deberse al azar, o, para ser más precisos, era demasiado probable que fuese un resultado del azar para considerarse significativa. En contraste, el desempeño promedio de Minnesota fue significativamente mayor que el de Maine.

■ **Figura 7.5.** Comparación de los resultados promedio de los estados en matemáticas de octavo grado, NAEP, adaptado de James S. Braswell *et al.*, *The Nation's Report Card: Mathematics 2000* (Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, Agosto de 2001). Figura 2.9

Minnesota (MN)	Montana (MT)	Kansas (KS)	Maine (ME)	Vermont (VT)	Massachusetts (MA)	North Dakota (ND)	Indiana (IN)	Ohio (OH)	Connecticut (CT)	Oregon (OR)	Nebraska (NE)	North Carolina (NC)	Michigan (MI)	DoDEA/DoDDS (DI)
MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN	MN
MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT	MT
KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS	KS
ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME	ME
VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT	VT
MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA	MA
ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND
IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN	IN
OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH	OH
CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT	CT
OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR	OR
NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE	NE
NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC	NC
MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI	MI
DI	DI	DI	DI	DI	DI	DI	DI	DI	DI	DI	DI	DI	DI	DI

Esto no indica que los promedios de Montana y Minnesota en verdad sean los mismos. Si se examinara a muchos más niños en ambos estados, posiblemente se encontraría que el desempeño de los estudiantes de Minnesota de hecho es un poco mejor, o viceversa. Más bien, sólo significa que, dado el error de esos datos, no puede tenerse mucha confianza en que la diferencia observada en los resultados es real. La probabilidad es tan alta que puede haber sido producto del error de muestreo. De igual modo, esos resultados no necesariamente significan que el desempeño de Maine habría sido de verdad inferior si se hubiese

examinado a todo mundo y la prueba no tuviese error de medición. Más bien indica que el promedio de Maine es lo bastante inferior al de Montana para que no sea probable que la diferencia haya resultado del azar.

El segundo error frecuente es que los hallazgos estadísticamente significativos son "importantes" o "esenciales" y que los resultados no significativos carecen de importancia. Esta conclusión nunca es fiable. Por ejemplo, una diferencia entre dos estados que sea demasiado pequeña para tener alguna relevancia práctica puede sin embargo ser estadísticamente significativa; y al contrario, un hallazgo de fundamental importancia puede resultar no significativo en un caso particular. Una de las varias razones para ello es que la significancia estadística depende del tamaño de la muestra así como del tamaño del resultado en cuestión. Una muestra grande da lugar a menos error de muestreo, lo que a su vez suscita un nivel más alto de confianza o de significancia estadística. De modo que una diferencia fundamentalmente pequeña puede ser estadísticamente significativa si la muestra es lo bastante grande y una diferencia considerablemente grande puede ser estadísticamente no significativa si la muestra es lo bastante pequeña. La significancia estadística sólo nos dice que es improbable que un determinado resultado haya surgido por azar.

La manera adecuada de usar la información acerca de la significancia estadística es, entonces, tratarla como una indicación de cuánta confianza puede tenerse en los resultados. Si sospecha que un hallazgo no significativo (por ejemplo, la diferencia entre dos escuelas o distritos) no es cuestión del azar a pesar de la falta de significancia estadística en un conjunto de datos, una opción sería buscar otros datos que aborden la misma cuestión –otros datos sobre el desempeño en el mismo año o los resultados de años posteriores.

Respuesta al error cuando se usan los resultados de las pruebas

La ubicuidad de los errores de medición y de muestreo no es razón para renunciar a la evaluación, pero sí señala la necesidad de usar los resultados de las pruebas con cautela y no tratarlas como indicadores únicos y perfectamente confiables del conocimiento y las habilidades de los alumnos. Sin importar lo valiosa que pueda ser, un resultado sólo representa a una única muestra del dominio –recuerde la prueba de vocabulario– y una sola ocasión de medición. El error resultante es un motivo por el que en el campo de la medición es axiomático que, en la medida de lo posible, las decisiones importantes no deben basarse en el resultado de una sola prueba, un axioma que suele ignorarse en la práctica, aunque está expresado claramente en los *Estándares para la Evaluación Educativa y Psicológica* publicados por las asociaciones profesionales más importantes.[10] Usar los resultados junto con otra información e ignorar las pequeñas diferencias (como recomienda el College Board) son respuestas razonables al error de medición. Cuando se emplean puntos de corte, una respuesta común al error de medición es brindar a los estudiantes que reprueban el examen una segunda oportunidad de presentarlo, de modo que se disminuya la probabilidad de que los estudiantes reprueben debido sólo al error de medición. Cuando se examinan resultados agregados, como un reporte del porcentaje de alumnos de las escuelas de un estado que alcanzan un estándar de competencia, algunos de los efectos del error de medición pueden disminuirse, pero se tiene el problema adicional del error de muestreo y la incertidumbre que produce. Tocaré de nuevo esos problemas al final del libro, cuando analice algunos usos razonables de los resultados de las pruebas. ■

Capítulo 8
Informe sobre desempeño: estándares y escalas

Al inicio de la película *This Is Spinal Tap*, uno de los protagonistas, el músico roquero Nigel Tufnel, muestra al cineasta Marty DiBergi un cuarto lleno de equipo y hablan sobre el amplificador favorito de Nigel:

Tufnel: Es especial porque, mira, los números llegan a 11, aquí en el tablero...

DiBergi: Y esos amplificadores llegan nada más a 10.

Tufnel: Exacto.

DiBergi: ¿O sea que suena más fuerte? ¿Suena un poco más fuerte?

Tufnel: Bueno, sí suena más fuerte, ¿verdad? No es 10. Aquí está el 10, en lo máximo. Si estás en la guitarra en 10, ¿a dónde vas? ¿A dónde?

DiBergi: No lo sé.

Tufnel: ¡Exacto! ¡A ninguna parte! Si necesitamos más para saltar el precipicio ¿sabes lo que hacemos?

DiBergi: Irse a 11.

Tufnel: Exacto. Exacto. Uno más fuerte.

DiBergi: ¿Por qué no hacen que suene más fuerte el 10? ¿Que sea el número más alto y hacen que suene un poco más fuerte?

Tufnel: [Pausa larga]: Este llega a 11.[1]

Este diálogo ilustra el problema principal del *escalamiento*, el proceso de asignar números o etiquetas a cualquier cosa que se esté midiendo; en el caso de Nigel, el volumen; en el nuestro, el logro de los estudiantes. Nigel no logra entender la distinción entre el volumen real y la escala (los números impresos en su amplificador) utilizada para representarlo. Tal vez también pensaría que en el lado estadounidense de la frontera hace más calor que en México porque los estadounidenses usan la escala de temperatura Fahrenheit y los mexicanos usan la escala Celsius. (Para quienes no están familiarizados con la escala de temperatura Celsius –o de centígrados– que se usa prácticamente en todos lados salvo en Estados Unidos, 95 grados Fahrenheit corresponden a 35 grados Celsius).

La gracia de la escena en *Spinal Tap* reside en lo evidente del malentendido de Nigel sobre la escala de volumen –él es descrito de manera repetida como poco perspicaz–, pero en el caso de la evaluación del logro el problema del escalamiento es muy complejo y los malos entendidos no son tan esporádicos ni tan aparentes. En el curso de los años se han construido diversas escalas para cumplir diferentes propósitos, y su conocimiento es esencial para poder entender los patrones de logro que las pruebas están diseñadas para describir.

Las diferentes escalas ideadas para describir el desempeño en las pruebas son de dos tipos. Un método consiste en elegir varios niveles de desempeño con base en el juicio, dividir la distribución del desempeño en esos puntos e informar luego del logro en términos de las categorías resultantes. Esto es lo que se hace en todos los sistemas actuales de evaluación basada en estándares: se utiliza el juicio para establecer niveles como "básico" y "competente" y se reporta el desempeño de los alumnos en términos de los intervalos resultantes en que caen sus resultados. La mayoría de los proponentes de este tipo de informe no consideran que un conjunto de niveles de desempeño sea una escala, pero lo es, aunque,

como veremos, no es una escala muy buena. La aproximación tradicional alternativa consiste en crear algún tipo de escala numérica para representar el rango de desempeño en la prueba. Existen muchas escalas de ese tipo, como las escalas numéricas arbitrarias (por ejemplo, la escala SAT que va de 200 a 800), los rangos percentiles y los equivalentes de grado).

Enfoco el trabajo de este capítulo de manera retrospectiva. Empiezo con los métodos más recientes de reporte en términos de estándares y paso después a las escalas más tradicionales. Lo hago de este modo porque los estándares son mucho más sencillos (o al menos lo parecen) y están muy de moda, por lo que algunos lectores quizá quieran desechar las escalas numéricas tradicionales que son más complejas. Sin embargo, los informes basados en estándares de cerca parecen menos atractivos que de lejos, y conocer sus limitaciones puede estimular a los lectores para entender las escalas tradicionales.

Estándares de desempeño

En Estados Unidos se ha generalizado el reporte basado en estándares del logro de los estudiantes. En algún momento, prácticamente a todos los padres se les dice que sus hijos están "avanzados", son "competentes", "parcialmente competentes" o algo por el estilo en materias como matemáticas y artes del lenguaje, y los periódicos están repletos de reportajes que informan sobre el porcentaje de estudiantes que alcanzan uno u otro de esos estándares de desempeño, más a menudo el denominado "competente".

Políticamente, este tipo de reporte ha ganado aceptación por la insatisfacción generalizada con las escuelas del país, la cual dio lugar a varios movimientos de reforma educativa en las décadas pasadas. Para muchos críticos de la educación pública, las formas

tradicionales de reportar el desempeño de los estudiantes eran inaceptables. Como veremos, las escalas tradicionales son puramente descriptivas y no reflejan ningún juicio inherente respecto del nivel de desempeño esperado. Además, muchas de las escalas tradicionales están referidas a normas y comparan el logro de cualquier estudiante o grupo con la distribución actual del desempeño. Muchos críticos argumentaron que la distribución real del desempeño era inaceptablemente baja y que el uso del reporte referido a normas implicaba una aceptación tácita de este *statu quo* indeseable. En su opinión, decir que un estudiante está por arriba del promedio crea una falsa sensación de éxito si el promedio en sí es inaceptablemente bajo. Además, los críticos del reporte referido a normas no quieren reportes meramente descriptivos en términos neutros. Quieren evaluar el desempeño comparándolo con metas explícitas.

En su opinión, la solución consistía en crear pruebas que reporten los resultados de los estudiantes en términos de estándares de desempeño. En la jerga que se difundió con rapidez y que sigue en uso, los *estándares de contenido* son afirmaciones de *qué* deben saber y poder hacer los estudiantes. Los defensores del reporte basado en estándares sostienen que su método dirigirá la atención general a la meta de mejorar el logro de los estudiantes. Muchos afirman que también creará un sistema en que todos los estudiantes puedan tener éxito, porque creen que, con el tiempo, casi todos pueden alcanzar un nivel superior al estándar. Pocos de los defensores de esta nueva aproximación al reporte de resultados tienen idea de la caja de Pandora que han abierto.

Las dos últimas re-autorizaciones de la Ley de Educación Primaria y Secundaria convirtieron al reporte basado en estándares en ley federal. La Ley para el Mejoramiento de las Escuelas Estadounidenses, una nueva autorización de la Ley de Educación Primaria y Secundaria de 1994, requería que los estados pusieran en práctica sistemas para los estándares de contenido y desempeño.

La Ley NCLB ordena que las pruebas estatales que se utilicen para satisfacer sus requisitos se reporten en términos de esos estándares, uno de los cuales deberá denominarse "competente", y estipula que las escuelas deben ser sancionadas de acuerdo con un sistema complejo para determinar si el porcentaje por arriba del criterio "competente" es adecuado en términos de los objetivos establecidos por la ley.

Este tipo de reporte muy pronto se popularizó entre los periodistas y entre los educadores, porque es sencillo, aparentemente claro y porque permite determinar si los estudiantes están a la altura de las expectativas. Sabemos lo que significa "competente", aunque no sepamos lo que significa una calificación o resultado de 156 en la escala. O más bien, la mayoría de la gente cree que sabe lo que significa "competente", aunque espero que después de algunas páginas más, usted ponga en tela de juicio la afirmación de que la mayoría de la gente en realidad lo entiende.

Además de ser sumamente popular entre los responsables de las políticas educativas y los periodistas, el informe basado en estándares ha recibido una bienvenida razonable en el campo de la medición. Lo cual refleja en parte el hecho de que la medición es, sobre todo, una profesión de servicio: si el gobierno insiste en que así debe informarse el desempeño en sus pruebas, la gente contratada para elaborar los exámenes estará obligada a hacerlo. Pero va más allá de eso; muchos miembros de la profesión han contribuido con entusiasmo al cambio hacia el reporte basado en estándares.

Lo anterior me resulta desconcertante porque el reporte basado en estándares presenta problemas serios. Debido a las maneras en que se establecen los estándares de desempeño, su significado es mucho menos claro de lo que cree la mayoría de la gente. Reportar de esta manera el desempeño en una prueba oculta información importante, exagera la importancia de algunos datos y puede proporcionar una visión considerablemente distorsionada

de diferencias y tendencias, además de que puede crear incentivos indeseables para los maestros.

Empecemos con la manera en que se establecen los estándares de desempeño. Para reportar el logro de los estudiantes en términos de unos cuantos estándares de desempeño es necesario decidir cuánto logro es suficiente para obtener la calificación o grado deseado. Una prueba típica de logro ofrece muchos resultados posibles. ¿En qué resultado o puntuación debe un estudiante pasar de "no competente" a "competente? ¿De simplemente "competente" a "avanzado"? ¿Cómo se toman en realidad las decisiones de dónde colocar los estándares de desempeño?

Los procedimientos que se usan más a menudo para establecer los estándares son complejos y misteriosos, y eso lleva a algunos de sus usuarios a suponer que el proceso es "científico" y, por ende, digno de confianza. La impresión que tengo cuando escucho a la gente describir los estándares de desempeño, es decir, a la gente que no conoce los detalles de cómo se establecen, es que casi siempre cree que hay alguna verdad subyacente acerca del desempeño, algún nivel real pero oculto del logro que constituye el ser "competente"; una verdad que de alguna manera es revelada por los complejos métodos utilizados para establecer los estándares. O, por lo menos, que los estándares establecidos rompen con claridad el continuo de desempeño en categorías inequívocas. Uno puede verlo, por ejemplo, en innumerables artículos periodísticos. Para tomar uno a la mano, un artículo publicado en el *Boston Globe* mientras yo escribía este capítulo declaraba: "Los estudiantes de Massachusetts van a la cabeza en las pruebas estandarizadas nacionales de matemáticas. Sin embargo, *menos de la mitad de los estudiantes del estado demuestran un sólido dominio de las matemáticas* en esas pruebas" (cursivas mías).[2] La reportera no ofreció más detalles, pero creo que se refería a la más reciente NAEP, que demostró que sólo 43 por ciento de los alumnos de octavo grado de Massachusetts

alcanzaron o superaron el "nivel de logro" (estándar de desempeño) competente.[3] Eso parece inequívoco ¿o no?

Sin embargo, una mirada más cercana al proceso nos muestra otra cosa. Se aplica el viejo chiste de que hay dos cosas que uno debería ver cómo se preparan: las leyes y las salsas. Yo añadiría los estándares de desempeño.

Aunque muchos métodos para establecer estándares se piensan con cuidado y se documentan y supervisan de manera escrupulosa, los resultados (qué nivel de desempeño se requiere para denominarse, digamos, "competente") siguen siendo cuestión de juicio. El juicio es algo confuso por la complejidad del proceso, pero no es reemplazado por algún criterio validado científicamente. Más aún, aunque los resultados del establecimiento de estándares parecen muy claros (un pequeño número de descripciones del desempeño expuestas por lo general de manera sencilla y aparentemente clara) son cualquier cosa menos eso. Un breve esbozo de dos de los métodos empleados con más frecuencia para establecer estándares en las evaluaciones de la educación primaria y media superior ilustrará este punto.[4]

Uno de los métodos más comunes, cuyas variantes se utilizan en la Evaluación Nacional del Progreso Educativo y en algunas evaluaciones estatales, es el llamado método *Angoff* o *Angoff modificado* (en honor al psicómetra William Angoff). En la versión del proceso Angoff modificado, puesto en práctica en la evaluación de matemáticas realizada por la NAEP en 2000, paneles de jueces comenzaron por considerar descripciones muy cortas, llamadas "definiciones políticas", de los estándares. Esas definiciones fueron las siguientes:

- *Básico.* Este nivel denota un dominio parcial del conocimiento y las habilidades requeridos que son fundamentales para el trabajo competente en cada grado.

- *Competente.* Este nivel representa un sólido desempeño académico para cada grado evaluado. Los estudiantes que alcanzan este nivel han demostrado competencia en contenidos temáticos que constituyen un reto. Incluyen el conocimiento del tema, su aplicación a situaciones reales y las habilidades analíticas apropiadas para el contenido temático.
- *Avanzado.* Este nivel significa un desempeño superior.[5]

Esas definiciones son muy vagas, como tendría que serlo cualquier descripción igualmente breve de los estándares de desempeño. ¿Qué tan bueno tiene que ser el desempeño para ser "superior"? ¿Qué contenido temático califica como un "reto"? ¿Qué es un "trabajo competente"? En el proceso de la NAEP, los propios panelistas añadieron algo de carne a esos huesos. Primero resolvieron la prueba y luego revisaron y elaboraron las definiciones de los estándares a partir de una lluvia de ideas acerca de cómo creían que debería ser el desempeño.

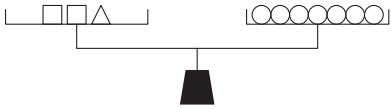
Luego, después de dos horas de entrenamiento en el método Angoff, calificaron cada uno de los reactivos. Para cada uno de los estándares (consideremos como ejemplo el estándar competente), se les pidió que imaginaran un grupo de estudiantes que lo rebasaran por muy escaso margen. Es decir, imaginaron a estudiantes que apenas calificaban como competentes. Luego tenían que calcular la probabilidad de que esos estudiantes imaginarios, escasamente competentes, respondieran cada reactivo de manera correcta.

Advierta que los panelistas no contaban hasta este punto con la ayuda de datos reales sobre el desempeño de los estudiantes. Tenían una definición del estándar elaborada a partir de la lluvia de ideas; tenían en su mente un grupo imaginario de estudiantes que apenas cumplían ese estándar impreciso y tenían, además, un conjunto de reactivos del examen. No disponían de ejemplos de alumnos que en realidad hubieran alcanzado el estándar. Hasta este

punto no contaban con información sobre la verdadera dificultad de los reactivos para los estudiantes, aunque sí sabían lo difícil que habían resultado para ellos. (La dificultad para los adultos puede ser una guía muy engañosa respecto de la dificultad para los estudiantes que aún están en la escuela, y esa diferencia puede ir en cualquier dirección, como suelen descubrir los padres de alumnos de preparatoria cuando tratan de desenterrar los detalles de sus antiguas clases de matemáticas para ayudar a sus hijos con las tareas.) Al carecer de datos, los maestros del grado en cuestión contaban con su propia experiencia como guía. Sin embargo, los paneles para el establecimiento de estándares suelen incluir a personas que no son docentes y que por lo general tienen poca o ninguna experiencia para poder avanzar.

Para concretar la dificultad de la tarea de los panelistas, inténtelo usted mismo con el reactivo de matemáticas que se muestra en la figura 8.1. Este reactivo se utilizó en la NAEP que se realizó en 2003 para cuarto grado. NAEP no publica los porcentajes de alumnos que superan por escaso margen cada uno de sus estándares y

Figura 8.1. Un reactivo de matemáticas tomado de la Evaluación Nacional del Progreso Educativo aplicada en 2003 para cuarto grado



13. Los objetos sobre la balanza presentada arriba logran que se equilibre de manera precisa. De acuerdo con esta balanza, si △ equilibra ○○○, ¿cuál de los siguientes entonces equilibra □?

A) ○

B) ○○

C) ○○○

D) ○○○○

que responden correctamente a los reactivos, por lo que vamos a cambiar un poco la tarea para que corresponda a los datos que proporcionan. La NAEP tiene tres estándares (básico, competente y avanzado) y, por ende, cuatro niveles de desempeño: por debajo de lo básico, arriba de lo básico pero por debajo de lo competente, por arriba de lo competente pero debajo de avanzado y por arriba de avanzado. Al hacer referencia a las definiciones anteriores para explicar el significado de los estándares, trate de calcular los porcentajes de cada uno de esos cuatro grupos que respondieron correctamente a este reactivo. Los porcentajes reales aparecen en el pie de página.*

Después de esta primera ronda de calificaciones, el procedimiento usado por la NAEP y por muchos otros programas de evaluación introduce algunos datos reales sobre el desempeño, llamados *datos de impacto*. Se proporciona a los panelistas el porcentaje real de alumnos que respondieron correctamente cada reactivo, el porcentaje de todos los estudiantes, no el de los alumnos en los grupos imaginados que superan apenas los estándares que todavía no se han establecido. Esto agrega a los estándares un elemento referido a normas porque los datos de impacto son de hecho datos normativos acerca del desempeño. La NAEP también proporciona a los panelistas datos sobre la variación en las calificaciones entre los panelistas.

Con esta información adicional, los panelistas calificaron por segunda vez los reactivos, luego recibieron otra ronda de retroalimentación y calificaron los reactivos por tercera vez. Al final de este proceso se empleó un procedimiento matemático para

* Por debajo de lo básico, 14 por ciento; básico, 29 por ciento; competente, 67 por ciento y avanzado, 92 por ciento. Advierta que la tarea de los panelistas era un poco más fácil porque habían elaborado definiciones de los estándares que usaban como referencia, pero no todos los procedimientos de establecimiento de estándares las proporcionan.

relacionar las calificaciones finales de los jueces con la escala de reporte original de la NAEP a fin de determinar hasta qué punto de la escala tenía que llegarse para alcanzar cada uno de los tres estándares. Por ejemplo, en octavo grado, el estándar básico se estableció en una calificación escalada de 262, el estándar competente se estableció en 299 y el avanzado en 333.

Otro método para establecer estándares que se ha vuelto popular con mucha rapidez (y que en la actualidad es el más común en los programas estatales de evaluación) es el llamado método del *marcador*. Para iniciar este proceso, el editor de la prueba ordena todos los reactivos en términos de su dificultad real. Igual que en el método Angoff, se proporciona a los panelistas definiciones cortas de los estándares y se les pide imaginar a los estudiantes que han alcanzado un determinado nivel de desempeño. Esos alumnos pueden ser los que alcanzaron apenas un estándar dado o todos los que lo superaron pero no llegaron al siguiente nivel. Vamos a utilizar de nuevo el ejemplo del estándar competente. Luego se les pide que revisen los reactivos en orden de dificultad y que se detengan en el que consideren que sería respondido correctamente por un porcentaje especificado de los estudiantes apenas competentes que imaginaron. Este porcentaje, llamado *probabilidad de respuesta*, suele establecerse en 67 por ciento, pero no hay una razón de peso para ello y los paneles han usado probabilidades de respuesta que fluctúan por lo menos entre 50 y 80 por ciento. Regresaré en breve a algunos datos acerca de los efectos de esta elección, pero por ahora vamos a asumir la probabilidad de respuesta de 67 por ciento. Como los reactivos se ordenaron por dificultad, los panelistas por fuerza están haciendo el juicio de que todos los reactivos anteriores fueron respondidos correctamente por más del 67 por ciento de los estudiantes imaginarios del grupo y que todos los reactivos posteriores por menos de ese 67 por ciento. En su raíz, las exigencias cognitivas impuestas a los

panelistas que usan este método son similares a las asignadas por el método Angoff: calcular los porcentajes correctos para grupos imaginarios de estudiantes sin contar con datos reales del desempeño. Sin embargo, también aquí puede ofrecerse retroalimentación a los panelistas, incluyendo los datos de impacto para el total de la población examinada y luego pueden repetir el proceso en varias ocasiones. Al final, el estadístico de dificultad que se utiliza para ordenar los reactivos ofrece un vínculo con la calificación escalada que luego se considerará equivalente al estándar competente.

Mis escuetas descripciones no hacen justicia a esos métodos. Se ha invertido mucho razonamiento y esfuerzo para perfeccionarlos. Pero esos esbozos son suficientes para mostrar que la posición cardinal del juicio en el establecimiento de estándares y que la base para esos juicios (en el caso de esos dos métodos comunes, estimaciones del desempeño del reactivo en grupos imaginarios de estudiantes) no inspira mucha confianza. Cualesquiera que sean las ventajas y desventajas de esos métodos, no son el medio para descubrir alguna "verdad" o estándar objetivo que está a la espera de ser descubierto.

Esto hace que los estándares resultantes sean mucho menos convincentes de lo que mucha gente cree, pero de ninguna manera significa que carecen de valor, y de hecho existe un antiguo debate entre los expertos en medición acerca de su utilidad. Un elemento de este debate es la discusión acerca de si los estándares son arbitrarios y caprichosos. Dos importantes expertos en medición –Jim Popham y Ron Hambleton– han argumentado por separado que "arbitrario" tiene un significado positivo que indica un uso apropiado del juicio y un significado negativo que sugiere capricho. Cada uno sostiene que, cuando el establecimiento de estándares se realiza con cuidado, es arbitrario en el sentido positivo, pero no caprichoso.[6]

Hasta este punto el argumento es correcto, pero pasa por alto el planteamiento principal. La mayoría de los procedimientos de establecimiento de estándares se realizan con mucho cuidado y pocos afirmarían que son caprichosos. Sin embargo, eso no implica que no deberíamos preocuparnos por su arbitrariedad. El problema es si los estándares, a pesar de su arbitrariedad, proporcionan información clara y útil, no engañosa.

Una base para juzgar si los estándares están a la altura a este respecto es si los resultados son lo bastante sólidos para que los usuarios de los resultados o calificaciones puedan confiar en ellos. Por ejemplo, vimos antes que sólo 43 por ciento de los estudiantes de octavo grado de Massachusetts alcanzaron el nivel competente en la más reciente Evaluación Nacional del Progreso Educativo. Este hallazgo pretende decir a la gente algo claro acerca del desempeño de los estudiantes del estado, y es claro que así lo tomó la reportera del *Globe*: concluyó, con tristeza, que menos de la mitad de los estudiantes "demostraron un sólido dominio" de las matemáticas. Pero ¿qué pasaría si los resultados reportados fuesen sumamente sensibles a los detalles de la manera en que se establecieron los estándares, detalles que son del todo irrelevantes para las conclusiones que los lectores están desprendiendo de los resultados? ¿Qué sucedería si otro método hubiese dado a los lectores del *Globe* la noticia de que "75 por ciento [para sacar un número del sombrero] de los estudiantes del estado demostraron un dominio sólido de las matemáticas"?

Por desgracia, este ejemplo, aunque hipotético, es realista: los resultados del establecimiento de estándares por lo general no son sólidos, y eso pone en duda la interpretación del desempeño reportado en términos de esos estándares. ¿Deberían los lectores del *Globe* ser más o menos pesimistas acerca de los alumnos de octavo grado dependiendo del método elegido? Y al ser enfrentados con una declaración tan definitiva como la citada arriba,

¿cuántos lectores del periódico matutino tendrían alguna idea de que los hallazgos dependen no sólo del logro de los estudiantes sino también de la elección –que para todo propósito práctico es arbitraria– de los métodos utilizados para establecer estándares?

Es probable que este planteamiento les siente mal a algunos lectores, por lo que un poco de evidencia empírica puede facilitar su aceptación. En 1989, Richard Jaeger, que sin lugar a dudas era uno de los más importantes expertos en el establecimiento de estándares en el mundo, publicó una revisión exhaustiva que demostró que los resultados de ese proceso por lo general son desiguales entre los métodos. Revisó 32 comparaciones publicadas y calculó la proporción de los porcentajes de estudiantes que métodos distintos de establecimiento de estándares clasificaban como reprobados. En el caso común (la mediana) el método más estricto para fijar los criterios clasificó como reprobados a una y media veces más estudiantes que el método más indulgente, y en algunos estudios la proporción observada fue mucho mayor.[7]

Investigaciones más recientes no dan mucha base para el optimismo. Por ejemplo, algunos estudios han demostrado que los métodos en que los jueces evalúan un reactivo a la vez, como sucede en el método Angoff, arrojan resultados que son incongruentes con los obtenidos cuando se evalúa un conjunto de trabajos reales de un estudiante, como una parte o la totalidad de un examen resuelto. Peor aún, elegir un método particular para establecer estándares, por la razón que sea, no significa que se tenga la victoria asegurada, ya que son cada vez más las investigaciones que indican que los detalles de la manera en que se pone en práctica ese método (una vez más, detalles que por lo general son irrelevantes para las conclusiones que la gente desprende de los resultados) pueden ocasionar grandes variaciones en los estándares de desempeño. Por ejemplo, paneles de jueces constituidos de manera diferente producen a menudo estándares muy distintos; además, los

jueces tienden a elevar o disminuir sus estándares dependiendo de la mezcla de formatos de los reactivos (de opción múltiple y de respuesta construida). También se ha encontrado que los jueces subestiman la dificultad de los reactivos difíciles y sobreestiman la dificultad de los reactivos sencillos, lo que puede llevarlos a establecer estándares más elevados cuando los reactivos que evalúan son más difíciles.[8] Cambiar la probabilidad de respuesta utilizada con el método del marcador –una elección arbitraria– puede tener efectos notables en la ubicación de los estándares.[9]

Una comparación del estado actual de los estándares de desempeño hace más evidente la arbitrariedad de todos ellos. Como han indicado varios comentaristas, los porcentajes de los alumnos que alcanzan un determinado umbral –digamos, el estándar competente– muestran variaciones espectaculares entre los estados. Uno esperaría que esta variación representara diferencias reales en el logro, pero es claro que no es así. Por ejemplo, un artículo reciente señalaba que el porcentaje de estudiantes que alcanza el criterio competente en lectura de cuarto grado según los estándares estatales es de 81 por ciento en Massachusetts, 83 por ciento en Alabama y 53 por ciento en Maine. Para cualquiera que esté familiarizado con los resultados obtenidos en las pruebas, esos resultados deben elevar una señal de alerta porque los estados de Nueva Inglaterra y de la región norcentral por lo general superan por gran margen a los estados del sur profundo. De hecho, los resultados más recientes de la NAEP ordenan a esos tres estados como uno esperaría: el porcentaje de estudiantes que alcanzan el nivel competente de dicha evaluación en Massachusetts duplica al de Alabama (44 contra 22 por ciento). En matemáticas de octavo grado las incoherencias son aún más sorprendentes. Un trabajo reciente preguntaba si es creíble que los porcentajes de quienes alcanzan el nivel competente sean de 63 por ciento en Alabama, 53 por ciento en Mississippi y 16 por ciento en Missouri. La NAEP confirma que no lo es.

Aunque Missouri se encuentra en la misma región, la puntuación que obtiene en la NAEP es considerablemente más alta y, de acuerdo con ese barómetro, los porcentajes de competencia son de 15 por ciento en Alabama, 13 por ciento en Mississippi y 26 por ciento en Missouri.[10] ¿Qué debe creer entonces un preocupado ciudadano de Alabama? ¿Son "competentes" casi todos los estudiantes del estado o casi ninguno? Un detallado análisis estadístico de los estándares estatales confirma que no son hechos casuales: los porcentajes considerados competentes en buena parte no tienen relación con los verdaderos niveles de logro de los estudiantes del estado.[11] Un estudio reciente comparó los estándares establecidos por tres pruebas de logro con normas nacionales con los estándares establecidos para la NAEP y encontró incongruencias igualmente dramáticas. Por ejemplo, la NAEP realizada en el 2000 clasificó como "competente" o "avanzado" a 17 por ciento de los estudiantes de doceavo grado; los porcentajes de las otras tres pruebas nacionales fluctuaban entre 5 y 30 por ciento.[12]

También es común que los estándares de desempeño sean incongruentes entre los grados o entre las materias de un grado. En la mayoría de los métodos usados para establecer estándares, un determinado panel de jueces considera sólo una materia y un grado. Por lo regular, no hay nada en el proceso que vincule los esfuerzos de los paneles de una materia o un grado al siguiente. Aunque en un programa de evaluación lo común es utilizar el mismo proceso para todas las materias y grados, los estándares de desempeño resultantes suelen diferir de manera considerable y, en ocasiones, dramática, entre grados o materias. Algunos estados están experimentando con algunos métodos para reducir esas incongruencias, pero todavía deben ser puestos a prueba y su impacto sigue siendo poco claro.

Para ser justos, la arbitrariedad de los estándares de desempeño no tiene por qué volverlos inútiles. Puede ser muy práctico contar

con una declaración formal de las expectativas; e incluso un estándar que al inicio es arbitrario con el tiempo puede adquirir significado gracias a la experiencia. Un resultado de 700 en la escala verbal de la prueba SAT al principio era sólo un número elegido de manera arbitraria, pero con la experiencia adquirió significado (aunque uno referido a normas): los estudiantes de preparatoria, los maestros, los padres y los encargados del departamento de admisiones saben que es una puntuación muy alta, lo bastante buena para mantener a un estudiante en la competencia para ingresar incluso a universidades muy selectivas. De igual manera, un estándar arbitrario calificado como "competente" con el tiempo puede adquirir significado a medida que la gente aprende qué nivel de trabajo requiere y qué estudiantes lo alcanzan.[13]

Sin embargo, este significado adquirido sólo puede funcionar hasta cierto punto. Las etiquetas elegidas para los estándares de desempeño, como "competente", tienen sus propios significados que son independientes de su uso con los estándares, y es claro que influyen en la manera en que la gente interpreta los resultados que se le presentan. La cita anterior del *Boston Globe* es un ejemplo: tener un "sólido dominio" de las matemáticas es un bonito sinónimo de "competente", ¿no es así? No obstante, dichas inferencias por lo general no están justificadas y es claro que en ocasiones son engañosas. Por ejemplo, el nivel y la descripción de los estándares usados en una de las encuestas del gobierno federal sobre la alfabetización de los adultos llevaron a muchas personas a inferir, sin razón, tasas elevadas de analfabetismo entre los adultos estadounidenses.[14] Las inconsistencias en los estándares entre materias y grados a menudo dan lugar a conclusiones como "nuestras escuelas son mucho más eficaces en inglés que en matemáticas" o "nuestras escuelas primarias son menos eficaces que las escuelas secundarias", cuando es posible que lo único que los datos reflejan sean diferencias en el establecimiento de estándares.

Independientemente de cómo se interpreten, los estándares impuestos ahora por los estados tienen un grave impacto práctico. Decisiones como si algunas escuelas están "fallando" según los términos de la ley NCLB y, en algunos estados, si un estudiante podrá obtener el diploma de preparatoria, pueden variar dependiendo de aspectos irrelevantes de los métodos usados para establecer los estándares de desempeño.

Incluso si se dejan de lado todas esas inconsistencias, el informe basado en estándares tiene un grave inconveniente: oculta una gran cantidad de información acerca de las variaciones en el desempeño del estudiante. Esto no es consecuencia de la naturaleza sentenciosa de los estándares, sino más bien de lo burdo de la escala resultante. Como se describió antes, la mayoría de los sistemas basados en estándares tienen tres o cuatro estándares de desempeño que crean cuatro o cinco rangos o categorías para reportar el desempeño. La información sobre las diferencias entre estudiantes *dentro* de cualquiera de esos rangos no se registra, y esas diferencias inadvertidas pueden ser muy grandes.

Como resultado, mejoras sustanciales pueden pasar inadvertidas mientras que ganancias triviales pueden parecer grandes. Por ejemplo, supongamos que un estado tiene estándares similares a los de la NAEP descritos arriba, en que una puntuación de 262 corresponde al nivel básico y se necesita 299 para alcanzar el nivel competente. Considere ahora una escuela que empieza con una gran cantidad de estudiantes en el extremo bajo del rango básico, digamos entre 262 y 275. A fuerza de trabajo duro y de la evaluación cuidadosa de los métodos de enseñanza, esta escuela logra llevar a la mayoría de sus alumnos a niveles justo por debajo del corte de 299 para el nivel competente. Considere ahora una segunda escuela que aunque no puede lograr mejoras de esta dimensión, se las arregla para empujar unos cuantos puntos a algunos estudiantes que están justo por debajo del nivel de "competente",

apenas lo suficiente para hacerlos pasar el estándar. La primera escuela ha logrado una mejora considerablemente mayor que la segunda, pero en términos del estadístico del porcentaje de alumnos que alcanzan el nivel de "competente" –el eje de la rendición de cuentas según la NCLB– parecerá que la primera escuela no logró ninguna ganancia en absoluto y que la segunda escuela hizo un progreso considerable.

Aunque este ejemplo puede parecer artificial, el problema es real y muchos maestros hablan abiertamente de los perversos incentivos que crean para ellos esas distorsiones. En un sistema de rendición de cuentas que se enfoca en el reporte basado en estándares, los maestros tienen un incentivo para concentrar sus esfuerzos en los estudiantes que están cerca del punto de corte entre estándares, ya que sólo se registrarán los cambios en el desempeño de esos estudiantes. Existe incluso un término común para identificarlos: son "los chicos de la burbuja". Según la NCLB, el incentivo es todavía más simple: enfocarse en los estudiantes que pueden llevarse al punto de corte competente e ignorar incluso los otros estándares de desempeño porque esos no importan para los propósitos de rendición de cuentas de la NCLB.

Un problema final inherente al reporte en términos de estándares es que dicho reporte puede distorsionar las comparaciones de las tendencias mostradas por diferentes grupos, como los integrantes de grupos minoritarios y mayoritarios o de los estudiantes con y sin discapacidades. Por ejemplo, hace unos cuantos años, una reportera del *Boston Globe* me llamó justo después de que se publicaron los resultados de ese año de la evaluación estatal del MCAS. La periodista afirmaba que los resultados demostraban que los estudiantes afroamericanos constituían una proporción creciente de los alumnos reprobados, es decir, de quienes no lograban alcanzar los estándares de desempeño relevantes. La pregunta que me hizo fue la siguiente: "¿No significa eso que el sistema les

está fallando a los estudiantes afroamericanos y que ellos se están quedando cada vez más rezagados?".

Para su evidente molestia, le dije que no tenía idea y que necesitaría otros datos para responder a su pregunta. Le expliqué que el problema es que cuando el desempeño se reporta en términos de estándares, las comparaciones de las tendencias de desempeño entre dos grupos que empiezan en distintos niveles (como los blancos y los afroamericanos en Boston) casi siempre son engañosas. Para este propósito se utilizan dos estadísticos diferentes. Uno es el que ella utilizó: la composición del grupo que no logra (o que consigue) alcanzar el estándar. Es más probable encontrar ese estadístico en discusiones acerca de la equidad en la admisión a la universidad, como en la pregunta ¿qué fracción de los estudiantes admitidos proviene de los grupos minoritarios que suelen obtener bajos resultados? El denominador de esta fracción es el número de estudiantes del grupo rechazado o aceptado. El segundo estadístico, mucho más común en los reportes de los resultados de educación primaria a media superior, es el porcentaje de cada grupo que alcanzó (o no logró alcanzar) un estándar, como el porcentaje de blancos o de afroamericanos que obtuvieron el nivel competente. En este caso, el denominador de la fracción es el número de estudiantes en el grupo en cuestión, como el de los afroamericanos. Los dos estadísticos son problemáticos, y por la misma razón.

Considere una comparación hipotética entre blancos y afroamericanos. Suponga que la diferencia entre las medias de los dos grupos es grande, que también lo es la variabilidad de los resultados dentro de cada grupo y que la mayoría de los estudiantes de cada grupo obtienen puntuaciones relativamente cercanas al promedio de su propio grupo. Eso es justo lo que suele encontrarse.

Imagine ahora algo poco común: suponga que cada individuo, sin importar el grupo al que pertenezca, mejora *exactamente* la

misma cantidad. Esto implicaría un progreso idéntico para ambos grupos; cada individuo hace el mismo progreso y toda la distribución de resultados de los afroamericanos ascendería al mismo ritmo que la distribución de los resultados de los blancos. En condiciones ideales, querríamos un resumen de las ganancias de los dos grupos de este ejemplo hipotético para demostrar este progreso igual. Y si le presentaran las tendencias en términos de los cambios en el promedio o la mediana de los resultados de los dos grupos, vería de hecho un progreso idéntico. Pero eso no sucedería si usara cualquiera de los dos tipos de estadísticos basados en estándares. Esos mostrarán cambios diferentes en ambos grupos y la naturaleza de la diferencia aparente dependerá de dónde se encuentre el estándar en relación con las dos distribuciones de resultados.

Por lo tanto, le dije a la impaciente reportera, la diferencia entre grupos que mencionó no es una medida directa de si los afroamericanos se están rezagando mucho de los blancos en las escuelas de Boston. Podría ser el caso, pero también es posible que no. Se lo podría haber dicho con los sencillos estadísticos que solían reportarse, como el promedio de los resultados escalados. Pero los reportes de la brecha en el logro ahora son dominados por las diferencias entre grupos en términos de estadísticos basados en estándares; en particular, cambios en el porcentaje que supera el estándar de competente. ¿Cuántos lectores de un reportaje que muestra diferentes incrementos en los porcentajes que alcanzan dicho estándar se darían cuenta de que eso no necesariamente indica que un grupo está rezagado o está al parejo del otro grupo?

Dada la debilidad del reporte basado en estándares, vale la pena regresar a las razones por las que sus defensores desdeñan con tanta frecuencia el reporte referido a normas. ¿Qué tan práctico es evitar el reporte referido a normas? ¿En realidad es cierto, como afirman algunos defensores del reporte basado en estándares, que

el reporte referido a normas no dice nada acerca de si el nivel de desempeño de los estudiantes es aceptable?

Los datos normativos a menudo se deslizan sigilosamente al escenario de estándares. En ocasiones esto sucede durante el proceso inicial de establecimiento de estándares, como cuando se entrega a los panelistas los datos de impacto. Otras veces ocurre después del hecho, cuando los políticos deciden que los estándares resultantes del proceso son inaceptables o poco razonables. A pesar de esta modesta dependencia de los datos normativos, los estándares se establecen a veces en niveles que, según los datos normativos, son poco razonables. Por ejemplo, al ser enfrentados con las enormes incoherencias en los niveles en que se establecen los estándares entre los estados, algunos críticos sostienen que todos los estados deberían usar estándares tan exigentes como el estándar de competencia establecido por la NAEP. Pero en el caso de matemáticas de octavo grado, ese estándar es tan elevado que no lograría alcanzarlo aproximadamente un tercio de los estudiantes de Japón y Corea, que se encuentran entre los países con mayores puntuaciones del mundo;[15] y como vimos en el capítulo 5, la brecha entre esos países y Estados Unidos es muy grande. Establecer, como objetivo a corto plazo, para todos los estudiantes estadounidenses un nivel de desempeño que no puede alcanzar la tercera parte de los estudiantes de Japón y Corea parece, en el mejor de los casos, poco realista.

En la práctica, todavía es frecuente el uso de reportes referidos a normas, aunque en ocasiones sin conocimiento, incluso de sus críticos. Un ejemplo serían los multicitados resultados de la NAEP que comparan a los estados en términos de los porcentajes de alumnos que alcanzan un determinado estándar, como el de competente. Podemos averiguar, por ejemplo, que en el año 2000, 22 por ciento de los alumnos de cuarto grado de Maryland alcanzaron o superaron el estándar competente. Basado en estándares

¿verdad? Pero aquí está la cuestión: ¿Cómo se sabe qué hacer con este resultado? ¿Veintidós por ciento es mucho o poco? Una manera de saberlo es comparar este porcentaje con los porcentajes de otros estados. La NAEP presenta juntos los porcentajes de todos los estados, ordenados del mayor al menor, lo cual resulta muy conveniente (¿se acuerda usted de la gráfica de pantimedia del capítulo 7?). Estas gráficas reportan el desempeño en términos de normas estatales, es decir, por comparación con una distribución del desempeño de los estados expresado en términos de los porcentajes por arriba del nivel competente. De este modo, la NAEP basa sus reportes tanto en estándares como en normas, y proporciona, además, otras *normas estatales* basadas en el promedio de los resultados escalados de los estados. Como se mencionó en el capítulo 5, las comparaciones internacionales como la prueba TIMSS también se sustentan en comparaciones normativas.

Lo que esos ejemplos ilustran es que es difícil evitar el reporte referido a normas porque es informativo y que es rutinario el uso de normas para evaluar las expectativas de desempeño. Por ejemplo, suponga que debe designar al nuevo entrenador del equipo de atletismo de una escuela secundaria. Un solicitante llega desbordante de entusiasmo y anuncia que su objetivo es lograr que en un año la mitad de los corredores de distancia recorran una milla en tres minutos y medio. ¿Qué haría usted? Lo mandaría a paseo porque es completamente incompetente o es un mentiroso. (Hasta donde sé, sólo un estudiante de secundaria ha corrido una milla en menos de cuatro minutos, y su tiempo estaba apenas abajo de cuatro. Ni siquiera los mejores corredores de distancia adultos de todo el mundo se acercan a recorrer una milla en tres minutos y medio.) En otras palabras, usted confía en la información referida a normas para saber que el nivel de desempeño que le promete es absurdo. Este ejemplo es inventado, pero el hecho es que en todos los aspectos de la vida hacemos uso constante de la información

normativa: para evaluar el kilometraje de los carros por litro de gasolina, para decidir si una compra es demasiado cara, etcétera. La evaluación a través de exámenes no es diferente.

Tal vez en respuesta a esto, algunos estados han empezado a añadir información normativa a sus reportes basados en estándares. Un buen ejemplo es el caso de Massachusetts, que utiliza los reportes basados en estándares como método principal para presentar los resultados de su evaluación MCAS. Pero los simples porcentajes por arriba de las diversos puntos de corte son insuficientes para que los educadores evalúen el desempeño de sus alumnos. Por lo tanto, el Departamento de Educación de Massachusetts agregó información normativa a sus reportes. Una parte de esa información consiste en comparaciones normativas de estadísticos basados en estándares: datos de cómo se comparan los porcentajes de una escuela por arriba de un estándar con el distrito y el estado como un todo. Massachusetts hace lo mismo con los porcentajes de aciertos en los reactivos individuales de la prueba. El estado proporciona a los educadores y al público información referida a normas para ayudarlos a entender los datos del desempeño basado en estándares.

Dados todos los problemas que surgen cuando se informa el logro de los estudiantes en términos de unos cuantos estándares de desempeño, ¿qué deberíamos hacer? En un artículo reciente en que bosquejó algunas de las debilidades más graves del reporte basado en estándares, Robert Linn, de la Universidad de Colorado, sugirió distinguir los casos en que se utiliza una prueba como base para tomar decisiones binarias, arriba o abajo (por ejemplo, al establecer la calificación aprobatoria de una prueba escrita de manejo o al usar las pruebas como criterio mínimo para entregar la licencia o certificación profesional) de los casos en que no es necesario hacerlo. Propuso que en el último caso –que incluye la mayor parte de las evaluaciones del logro en las escuelas primarias y secundarias– sería

mejor evitar el reporte basado en estándares.[16] Coincido con esa idea aunque, como también señaló Linn, es poco probable que tengamos esta opción, es posible que el reporte basado en estándares permanezca con nosotros por algún tiempo. De ser así, será necesario complementarlo con algo más que no comparta las mismas debilidades, y eso nos lleva a las escalas tradicionales.

Escalas

Dejemos entonces a un lado los estándares de desempeño y empecemos desde cero. Supongamos que alguien ha aplicado una prueba a alumnos en su área. Digamos que la prueba está compuesta por 50 reactivos. ¿Cuál sería una forma útil de reportar los resultados?

El método más sencillo sería llevar la cuenta para cada alumno: el número de reactivos que respondió correctamente o (si hubiera variaciones entre reactivos en el crédito posible) el número total de puntos obtenidos. Para no confundir esas cuentas con la extensión de la prueba, podríamos convertirlas en simples porcentajes: el porcentaje de reactivos respondidos correctamente o el porcentaje de posibles créditos obtenidos. A eso le llamamos *puntajes crudos*.

Este es el tipo de calificación de pruebas con el que todos crecimos y tiene cierta utilidad. Después de cada examen que aplico en mis grupos, presento una gráfica que muestra la distribución de los resultados crudos. Esto proporciona a los alumnos información valiosa referida a normas: una comparación de su desempeño con el del resto del grupo.

Pero como mencioné en el capítulo 1, esos simples resultados crudos tienen graves limitaciones. La más importante es que si no sabemos qué tan difícil fue el examen, no podemos evaluar qué tan buena es una determinada calificación. Uno podría diseñar

exámenes lo suficientemente difíciles para que la mayoría de los estudiantes no obtuviese ningún crédito o lo bastante fáciles para que casi todos obtuvieran calificaciones casi perfectas, aunque su verdadero conocimiento del material del curso fuese idéntico en ambos casos. Por supuesto, no tengo motivos para hacer nada de eso y en la práctica trato de elaborar exámenes que tengan más o menos el mismo nivel de dificultad de un año al siguiente. Pero a menos que mis exámenes sean prácticamente idénticos, no puedo confiar en que su grado de dificultad es en efecto el mismo. Por supuesto, la probabilidad es aún menor en el caso de exámenes escritos por diferentes maestros. Por lo tanto, la convención común de asignar como calificación una letra por un porcentaje fijo (90 por ciento de créditos equivale a una A, o algo por el estilo) no resulta útil para los programas de evaluación a gran escala. Puede servir dentro del salón de clases si el maestro entiende bien la dificultad de sus tareas y exámenes, pero no permite hacer las comparaciones entre escuelas y a lo largo del tiempo que requieren las evaluaciones a gran escala.

Enfrentados con esta limitación de los resultados crudos, los psicómetras han desarrollado diversas escalas como sustitutas de las calificaciones crudas. Las escalas fueron diseñadas con propósitos diferentes y, por desgracia, pueden brindar visiones distintas del desempeño de los estudiantes. Para aclarar lo anterior es útil considerar dos escalas que no tienen nada que ver con la evaluación.

Consideremos primero las escalas de temperatura. Imagine, por el momento, que uno de sus amigos está pensando mudarse de la ciudad A a la ciudad B y dice: "En la ciudad A la temperatura desciende por la noche casi tanto como en la ciudad B. La diferencia entre el promedio de la temperatura alta por el día y el promedio de la temperatura baja por la noche es de alrededor de 9 grados". Quizá no conozca gente que en realidad se fije en ese tipo de cosas pero, en aras de la explicación, sígame la corriente.

Suponga ahora que su amigo es estadounidense y que ambas ciudades se localizan en Estados Unidos, por lo que él se refiere a 9 grados Fahrenheit.

Imagine ahora que Estados Unidos decide hacernos a todos un favor, seguir el ejemplo del resto del mundo y adoptar por fin el sistema métrico. ¿Qué sucedería con la conclusión de su amigo acerca de la similitud de ambas ciudades?

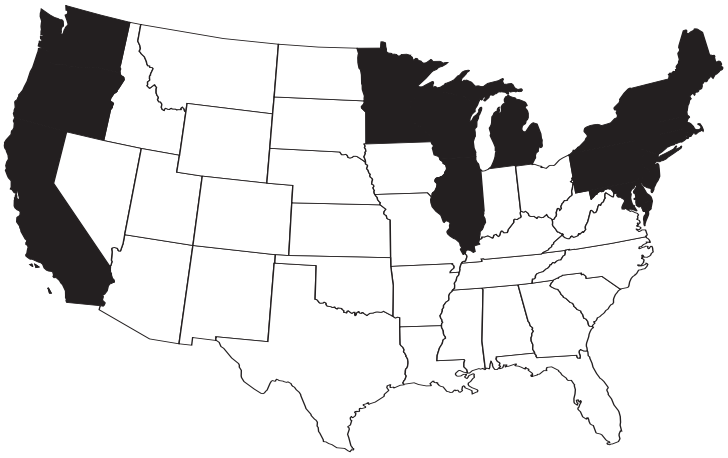
Nada, por supuesto. La diferencia de temperatura entre la temperatura alta diurna y la baja nocturna se expresaría como 5 grados Celsius en lugar de los 9 grados Fahrenheit, pero la diferencia de temperatura es la misma y no cambia la conclusión de que las dos ciudades son iguales a este respecto. Esto puede parecer obvio (aunque quizá no para Nigel Tufnel), pero no es así con todos los cambios de escala. En este ejemplo es cierto porque el cambio de Fahrenheit a Celsius es una *transformación lineal*, lo que significa que se consigue multiplicando por una constante y sumando otra. La conversión de Celsius a Fahrenheit requiere que se multiplique por 1.8 (porque los grados están más alejados en la escala Celsius) y se sume 32 (porque el valor de cero se encuentra en lugares distintos, en el punto de congelación del agua en el caso de la escala Celsius y en una temperatura insignificantemente menor en la escala Fahrenheit). Se le llama transformación lineal porque se efectúa aplicando una simple ecuación lineal en una variable de la forma $y = a + bx$. Cuando una transformación es lineal, cualquier par de diferencias que son del mismo tamaño en una escala (9 grados Fahrenheit en ambas ciudades) serán del mismo tamaño en la otra escala (5 grados Celsius en ambas ciudades).

La mayor parte de los cambios de escala que encontramos en la vida diaria (de gramos a libras, de dólares a euros, de litros a onzas, de pies cuadrados a yardas cuadradas) son transformaciones lineales, por lo que es fácil perder de vista el hecho de que no tienen que serlo. Considere la elección presidencial del 2004.

En los años recientes se ha vuelto común referirse a los estados o condados que votan por los republicanos como "rojos" y a los que votan por los demócratas como "azules". Después de la elección presidencial del 2004, los expertos empezaron a informarnos acerca de la enorme marea roja que había dejado a los votantes azules en lugares pequeños, aislados y algo extraños, como Massachusetts (mi estado) o California. El conocido mapa de cómo votaron los estados, que se reproduce en la figura 8.2, parecía respaldar su argumento: grandes franjas rojas (mostradas en blanco en la figura) aislaban a tres áreas azules relativamente pequeñas (aquí en negro).

Estaban equivocados, y no sólo porque atribuyeron persistencia a un resultado electoral que se desvaneció apenas dos años después. La figura 8.2 es en esencia engañosa porque usa la *escala*

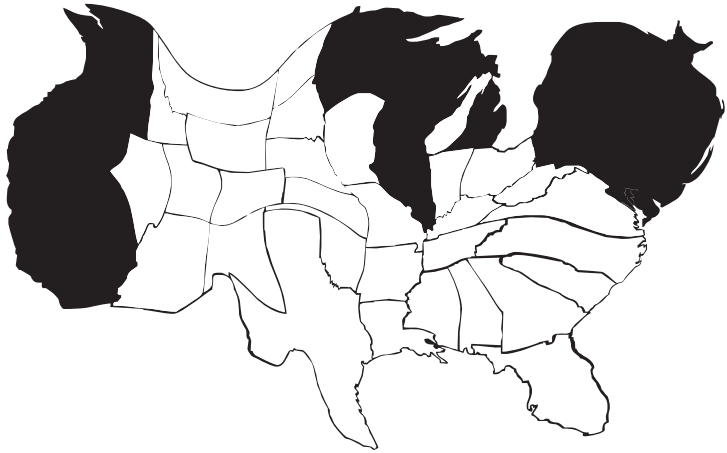
Figura 8.2. La elección presidencial del 2004, con los estados "azules" en negro y los estados "rojos" en blanco, escalados por área geográfica. Michael Gastner, Cosma Shalizi y Mark Newman, "Maps and Cartograms of the 2004 US Presidential Election Results", *http://www.personal.umich.edu/~mejn/election/* (consultado el 30 de diciembre de 2006)



equivocada; ordena a los estados en relación con la superficie en acres. Sin embargo, los acres no votan. Montana aparece como uno de los estados más grandes, pero en población es apenas un poco mayor que el área metropolitana de Albany, Nueva York, que ni siquiera es considerada una verdadera área metropolitana por muchos residentes del estado azul.

¿Cómo se vería entonces el mapa si se describiese a los votantes con una escala más razonable que clasifique a los estados en orden de población? El resultado, cortesía de dos físicos y un estadístico de la Universidad de Michigan, se muestra en la figura 8.3. Este mapa distorsiona muchísimo las dimensiones físicas y las ubicaciones de los estados, pero los ordena adecuadamente en orden de población.

Figura 8.3. La elección presidencial del 2004 con los estados "azules" en negro y los estados "rojos" en blanco, escalados por población. Michael Gastner, Cosma Shalizi y Mark Newman, "Maps and Cartograms of the 2004 US Presidential Election Results", *http://www.personal.umich.edu/~mejn/election/* (consultado el 30 de diciembre de 2006)



Montana ahora es mucho más pequeño que Massachusetts. Aunque las áreas azules todavía están aisladas, ya no parecen pequeñas, y no deberían puesto que su población total no es muy diferente a la del área roja.

Este es un ejemplo de una *transformación no lineal*, un cambio en la escala que no conserva las relaciones entre las observaciones, en este caso, entre los estados. Nueva Jersey se volvió mucho más grande, Montana mucho más pequeño y Kentucky no cambió mucho. Una comparación hecha con la primera escala (digamos, que Arizona y Nuevo México son más o menos iguales) no sería preservada por la transformación al segundo mapa porque Nuevo México es similar a Arizona en tamaño, pero su población es menor.

Por desgracia, muchas de las escalas usadas en la evaluación educativa son transformaciones no lineales de cada una. En su mayor parte, la elección de la escala no afecta las clasificaciones (los estudiantes que obtienen las puntuaciones más altas en una escala las obtienen también en la otra) pero sí cambia las *comparaciones*. Dos grupos que muestran el mismo progreso a lo largo del tiempo en una de esas escalas pueden no hacerlo en la otra.

Un tipo común de escala usa valores esencialmente arbitrarios. Hemos visto varias escalas de este tipo en los capítulos previos, incluyendo las que se utilizan para reportar los resultados en la prueba SAT, la ACT, la Evaluación Nacional del Progreso Educativo y la TIMSS. El hecho de que son escalas arbitrarias (en un sentido del término) se demuestra por lo diferente que ellas son entre sí. Por ejemplo, como sabemos, la escala de la prueba SAT (dentro de un área temática) va de 200 a 800, mientras que la de la prueba ACT va de 1 a 36. No hay razón de peso para esta diferencia; nadie podría argumentar que un estudiante que obtiene la puntuación más alta en la prueba SAT sabe 22 veces más que un estudiante que obtiene la puntuación más alta en la prueba ACT. Y la clasificación de los sustentantes no se modificaría si se cambiara la puntuación

ACT para ajustarla a la escala SAT, o si el College Board decidiera cambiar la escala de la prueba SAT a la escala ACT.

Hace algunos años un colega y yo presentamos un memorándum a un organismo gubernamental responsable de un importante programa de evaluación y nos referimos a la escala usada por el programa como arbitraria. Una integrante del personal del organismo se sintió agraviada y me reprendió diciendo que la escala no era arbitraria y que se había construido con mucho cuidado. Tenía razón en el hecho de que la escala se había elaborado con cuidado, pero nosotros teníamos razón en que era arbitraria.

El hecho de que esas escalas sean arbitrarias no implica que sean creaciones casuales, todo lo contrario. La elaboración de esas escalas conlleva muchos pasos, pero resulta útil la simplificación de considerar que tiene dos etapas. La primera, que es compleja y exigente si se hace bien (como sucede con los cuatro programas de evaluación que mencioné), da por resultado una escala provisional que posee características técnicamente útiles pero que usa una métrica inadecuada para el consumo público. En esta primera etapa se abordan casi todas las complicaciones de la elaboración de una escala, como relacionar la que se usará para los resultados más recientes con las que se emplearon en el pasado. Muchas veces, esta escala provisional se aproxima a la estandarizada que se describió en el capítulo 5, con una media cercana a cero y una desviación estándar (una medida de la dispersión de los resultados) próxima a uno. Como se mencionó en el capítulo 5, esas escalas suelen ser convenientes para los estadísticos y los científicos sociales, pero no son bien recibidas por el público porque asignan a los estudiantes resultados fraccionales y negativos. (Recuerde: ¿cuántos padres entenderían que a pesar de que su hijo respondió correctamente muchas preguntas recibe una calificación de cero o, peor aún, una calificación negativa?) La segunda etapa del escalamiento, que es más sencilla y a menudo arbitraria, evita esos

problemas transformando la escala provisional para hacerla más agradable. La forma más fácil de lograrlo es agregar un número arbitrario a todos los resultados (para obtener un promedio más alto) y luego multiplicarlos por otra constante arbitraria (para aumentar su dispersión). En el caso de las pruebas SAT y TIMSS, se estableció la escala para que tuviera una media de 500 y una desviación estándar de 100. En 2005, los resultados de matemáticas de cuarto grado de la Evaluación Nacional del Progreso Educativo tuvieron una media de 237 y una desviación estándar de 29. Cuando mi colega y yo calificamos de arbitraria la escala en cuestión nos referíamos sólo a esta segunda etapa; la empleada del organismo sólo pensaba en la primera etapa cuando nos replicó.

Sin importar el cuidado con que se hayan elaborado, las escalas formadas por números arbitrarios no parecen ser una manera muy útil de describir el logro de un estudiante. Pero esas escalas tienen algunas ventajas muy importantes sobre los resultados crudos y los estándares de desempeño.

La ventaja más importante de esos resultados arbitrarios escalados sobre los resultados crudos es que pueden hacerse comparables en el tiempo y entre las formas de la prueba. Con la misma facilidad con que el College Board decidió que la media de la prueba SAT fuese 500 podría haberla fijado en 320 o en 40, pero una vez que eligió 500, un resultado de 500 significa lo mismo sin importar cuándo se presenta la prueba, incluso si el resultado crudo varía un poco dependiendo de los reactivos incluidos en cada forma. Una ventaja relacionada de esas escalas sobre los resultados crudos es que sus propiedades (como las desviaciones estándar) son conocidas, de modo que si usted desea información más detallada, como la puntuación que se necesita para entrar al cuartil superior, es posible calcularla con facilidad. Existen excepciones (por ejemplo, las pruebas MCAS de Massachusetts están escaladas de una manera tan inusual que los resultados escalados

que se dan a conocer no proporcionan este tipo de información), pero son raras.

Usted podría replicar que en muchos programas de evaluación se ha logrado que los estándares de desempeño puedan compararse de un año al siguiente. De hecho, la ley NCLB y muchos de los programas estatales de rendición de cuentas que la precedieron dependen de esto: no tiene sentido responsabilizar a las escuelas de incrementar el porcentaje de alumnos que alcanzan un estándar de competente si la exigencia de ese estándar no permanece constante. Si la prueba se vuelve un poco más difícil o más fácil cada vez que se introducen nuevos reactivos, entonces el resultado crudo que se necesita para alcanzar el estándar debe cambiar para que éste permanezca fijo. Lo que no suele ser evidente a menos que se lean los reportes técnicos es que la clave para esta consistencia suelen ser los resultados escalados. Los resultados de la prueba se colocan sobre una escala; luego, los resultados escalados se relacionan estadísticamente para hacerlos comparables entre formas o años; finalmente, esos resultados escalados relacionados se usan para hacer comparables a los estándares. Los estándares de desempeño no ofrecen un medio de establecer la comparabilidad, simplemente se estratifican en la parte superior de la escala que proporciona esta oportunidad.

Muchas veces no se reporta la escala subyacente por el entusiasmo actual por los reportes basados en estándares. No hace mucho tuve una reunión con el director de un programa estatal de evaluación en que yo sostenía, en vano, que debía empezar a reportar los resultados escalados junto con los estándares de desempeño. Él contestó que eso significaría mucho trabajo para ellos. Señalé que ya disponían de una escala, la utilizaban cada año para vincular sus estándares con la prueba del año anterior. Todo lo que tenían que hacer era transformarla a una métrica más aceptable, que no era otra cosa que aritmética simple. El hecho de

que nunca lo hizo dice algo acerca del dominio del reporte basado en estándares.

Aunque en cierto sentido también son arbitrarios, esos resultados escalados tienen muchas ventajas sobre los estándares de desempeño. Las escalas son mucho más útiles que los estándares para describir toda la distribución del desempeño. Para dar un ejemplo sencillo, suponga que quiere seguir la trayectoria de las tendencias en el desempeño promedio de un grupo de estudiantes. Los estándares no permiten hacerlo de manera atinada; y a diferencia de los estándares de desempeño las escalas no crean distorsiones cuando se comparan tendencias entre diferentes grupos de estudiantes.

En algunos aspectos la arbitrariedad de los resultados escalados es menos problemática que la de los estándares de desempeño. Cuando se establecen dichos estándares, los niveles que se esperan de los estudiantes son arbitrarios y las etiquetas que se les asignan suelen acarrear significados previos, algunos de los cuales son injustificados. Modificar esas decisiones (seleccionar un método diferente que mueve un estándar hacia arriba o hacia abajo o asignar al estándar una etiqueta diferente) puede tener un efecto importante en la manera en que la gente interpreta el desempeño del estudiante, aun cuando el desempeño en sí puede no haber cambiado. Los resultados escalados no tienen esta carga. Pocas personas leen el reporte de un nuevo programa de evaluación con alguna idea previa de lo que significa una puntuación de 340. Para dar un ejemplo concreto, muchos lectores estarán familiarizados con la prueba PSAT (*Preliminary SAT/National Merit Scholarship Qualifyng Test*), una prueba que por diversas razones presentan muchos estudiantes de segundo y tercer año de preparatoria, en parte como preparación para presentar uno o dos años más tarde la prueba SAT. La escala usada con la PSAT es en esencia la escala de la prueba SAT pero dividida entre 10, por lo

que no deben confundirse –es decir, corre de 20 a 80 en lugar de hacerlo de 200 a 800. Todavía tengo que escuchar a personas que expresan sorpresa de que un estudiante que apenas obtuvo 70 en la prueba PSAT obtuvo alrededor de 700 en la SAT. Los números usados para las dos escalas no tienen significado previo que pueda desordenar la comprensión de la gente sobre los resultados de las pruebas.

Es obvio que esa fortaleza también es una debilidad: los números arbitrarios sin significado previo no son muy informativos. Es necesario hacer algo para asignarles significado. ¿Cómo va a saber un padre si una puntuación de 29 en la prueba ACT es buena o mala noticia?

Una vez más, la respuesta está en las normas. A menudo esta información normativa es informal y aproximada. Por ejemplo, la mayoría de los padres de estudiantes que piensan asistir a la universidad en estados en que domina la prueba SAT escuchan hablar constantemente de ella; leen artículos ocasionales en los periódicos acerca de las calificaciones SAT; quizá han tenido una conversación con los orientadores escolares acerca de los resultados de sus hijos y con frecuencia conocen los resultados de otros estudiantes. Es posible que también hayan recibido promociones de las empresas de preparación para las pruebas en que se mencionan los resultados. Gracias a esta exposición saben que una puntuación de 600 en matemáticas es moderadamente alta.

La NAEP es un ejemplo interesante porque los organismos responsables han luchado por años con el problema de añadir significado a los resultados escalados y han ideado diversos métodos para ello. Por supuesto, uno ha sido estratificar los estándares de desempeño ("niveles de logro") sobre la escala. Pero muchos de los métodos para dar significado a los resultados de la NAEP se basan en información normativa. Por ejemplo, los datos normativos son fundamentales para dar sentido a las muy publicitadas

comparaciones que hace la NAEP entre los estados. ¿Cómo va a saber el comisionado de educación de Minnesota si la puntuación escalada promedio del estado de 290 es buena o mala? Una forma de saberlo es comparar ese promedio con la distribución de promedios estatales, que se presentan en los reportes de la NAEP. La distribución muestra que 290 es bastante bueno, al menos hasta donde concierne al logro en Estados Unidos. La NAEP también brinda, para cada estado, la proporción de estudiantes que cae en cada uno de los cuatro "intervalos" de desempeño creados por los niveles de logro, es decir, una combinación de reporte basado en estándares y referido a normas.[17]

Los resultados escalados son acompañados, a menudo, por una segunda escala, explícitamente normativa, con más frecuencia los rangos percentiles (RP). Un rango percentil es simplemente el porcentaje de resultados que caen por debajo de un determinado nivel. Tomemos de nuevo como ejemplo la puntuación de 600 en la prueba SAT, los padres pueden pensar que es "moderadamente alta", pero una mirada a la Red revela información más precisa: en el 2005, una puntuación de 600 en matemáticas correspondió al percentil 75, es decir, 75 por ciento de los estudiantes que presentaron la prueba obtuvieron puntuaciones inferiores a 600.[18] Los rangos percentiles no tienen nada que ver con el porcentaje de reactivos que un alumno responde correctamente, son sólo una manera de plantear cómo le fue al estudiante en comparación con los demás. Muchos tipos de pruebas de admisión dan a los alumnos un rango percentil junto con una puntuación escalada.*

En este caso resulta relevante el ejemplo de la elección del 2004: la transformación de los resultados escalados a rangos percentiles

* Técnicamente, se denomina percentil a la calificación en sí (en este caso 600), mientras que el porcentaje (en este caso, 75) es el rango percentil. Sin embargo, en la práctica suele emplearse para ambos el término *percentil*.

no es lineal. En la mayoría de las pruebas, los resultados de muchos estudiantes se agrupan cerca del promedio y son muy pocos los que obtienen resultados muy altos o muy bajos. Podemos usar de nuevo la prueba SAT como ejemplo. Un estudiante cuya puntuación aumenta de 450 a 550 (de un poco por debajo del promedio a un poco por arriba) mostrará un incremento considerable en el rango percentil. Como son muchos los estudiantes apilados en esa región, rebasa a muchos de ellos cuando pasa de 450 a 550. Sin embargo, un estudiante que empieza en 650 y muestra la misma ganancia de 100 puntos en los resultados escalados, obtendrá un incremento mucho menor en el rango percentil porque son mucho menos los estudiantes en el rango de 650 a 750. De este modo, las dos escalas describen de manera diferente a los dos estudiantes, una sugiere la misma mejora y la otra una mejora diferente. ¿Cuál es correcta? Ambas lo son. Sólo que se refieren a cosas diferentes con "mejora".

Otra categoría de escalas usadas con algunas pruebas de logro suelen denominarse *puntajes estándar*. Estas difieren de las escalas arbitrarias ya mencionadas en que adoptan exactamente la misma forma de una prueba a otra. Se basan en la escala estandarizada que se explicó en el capítulo 5 (media de 0, desviación estándar de 1), pero se transforman para tener diferentes valores. Una de esas escalas se denomina *puntajes T*, la cual tiene una media de 50 y una desviación estándar de 10. Otra más, una peculiar innovación surgida de los requisitos de evaluación del Título I de la década de 1970 a que se hizo referencia en el capítulo 4, son los *equivalentes de la curva normal,* denominados simplemente como ECN, los cuales se transforman para tener una media de 50 y una desviación estándar un poco mayor a 21. Los puntajes estándar se proporcionan a menudo con las pruebas tradicionales de logro referidas a normas, pero por lo general no se utilizan con las pruebas más recientes diseñadas para estados específicos. Los educadores y padres de

familia todavía pueden encontrarlas en los distritos o estados que usan pruebas referidas a normas, pero es raro que las encuentre el público general.

Una última categoría son las *escalas de desarrollo*, que se diseñan específicamente para medir el crecimiento en el logro a medida que los estudiantes progresan en la escuela. Las más antiguas son los *equivalentes de grado* o EG. En el curso de las décadas pasadas, los equivalentes de grado perdieron el favor de la gente, lo cual es una pena porque son fáciles de entender y proporcionan una manera intuitivamente clara de pensar en el desarrollo de los niños. Un equivalente de grado es simplemente el desempeño típico –el desempeño del estudiante medio– en cualquier nivel escolar. Por lo general se presenta en términos de años y meses académicos (con 10 meses académicos por año escolar). De modo que un EG de 3.7 es el desempeño medio que exhiben en marzo los estudiantes de tercer grado en la prueba diseñada para ellos. Los equivalentes de grado informan si los estudiantes mantienen el paso de la norma del grupo. Por lo tanto, si una alumna de tercer grado obtiene un equivalente de grado de 4.7, eso significa que su desempeño en la prueba de tercer grado está muy por arriba del promedio, específicamente es comparable al desempeño que exhibiría en marzo el estudiante medio de cuarto grado en la prueba de tercer grado. En el capítulo 10 voy a utilizar la escala equivalente de grado para demostrar lo grave que puede ser el problema de la inflación de resultados o calificaciones cuando se hace a los educadores responsables de los resultados obtenidos en las pruebas.

A pesar de su utilidad, los equivalentes de grado tienen varios inconvenientes. Uno es que la tasa de crecimiento en un área temática no es constante a medida que los niños se hacen mayores. Por ejemplo, el niño típico adquiere las habilidades de lectura con mayor rapidez en los grados de primaria que más adelante. Por lo tanto, una ganancia de un equivalente de grado denota un

mayor crecimiento en los primeros grados que en los posteriores. Por ejemplo, si quiere saber si la tasa con que los estudiantes aprenden matemáticas se hace más lenta o se acelera cuando pasan a la secundaria, los equivalentes de grado no se lo pueden decir por su propia naturaleza. El estudiante promedio ganará un equivalente de grado por año pase lo que pase.

La última escala que voy a mencionar (la cual aparece con frecuencia) es otra escala de desarrollo que trata de evitar esta limitación de los equivalentes de grado. Se conoce por el nombre apropiado aunque engorroso de calificación estándar de desarrollo, que a menudo es abreviado de manera confusa como *calificación de desarrollo* o sólo calificación escalada. Es una escala numérica arbitraria, aparentemente parecida a las escalas utilizadas para la NAEP o la prueba SAT, pero con una diferencia importante: las calificaciones estándar de desarrollo están relacionadas entre los grados de una manera que se supone da a cualquier incremento específico en el desempeño el mismo significado en cada grado. Por ejemplo, suponga que un estudiante mostró una ganancia de 230 a 250 entre tercer y cuarto grados, y que otro aumentó de 245 a 265 entre cuarto y quinto grados. Si son calificaciones estándar de desarrollo, sus ganancias idénticas de 20 puntos significarían idealmente que ambos mejoraron su desempeño en la misma cantidad. A pesar del nombre incoherente y confuso de esas escalas, a menudo puede identificarlas comparando los números entre grados. Si los números entre grados son similares, la escala no es una calificación estándar de desarrollo, pero si se incrementan de un grado a otro (y los resultados no están dados en una escala equivalente de grado), es muy posible que lo sean.

En todos los campos, se denomina *escalas de intervalo* a las escalas que tienen esta propiedad, es decir, a las escalas en que cualquier diferencia dada tiene el mismo significado en diferentes niveles. La mayor parte de las escalas que usamos en la vida diaria

son escalas de intervalo. La longitud es obviamente una escala de intervalo –un metro más de cuerda tiene la misma longitud y le costará lo mismo en la ferretería, sin importar si ya sacó del carrete 10 o 20 metros.

Por desgracia, las calificaciones escaladas de desarrollo no pueden considerarse verdaderas escalas de intervalo. Rara vez es posible confirmar que una ganancia de 10 puntos significa lo mismo en diferentes grados, y a veces es claro que no es así. Crear semejante escala no es una meta del todo práctica. Para crear una verdadera escala de intervalo los estudiantes tienen que estar aprendiendo lo mismo a lo largo de todo el rango de grados en cuestión. Eso sería una suposición razonable si se estuviera comparando, digamos, la lectura en segundo y tercer grados, pero sería muy forzado cuando se compara matemáticas de tercer grado con matemáticas de séptimo grado. ¿Qué tanta facilidad con la multiplicación básica es equivalente a una determinada ganancia en preálgebra? Una regla práctica sensata consiste en tratar a esas escalas como aproximadas y mostrarse cada vez más escépticos a medida que aumenta el rango de grados que cubren.

Enfrentados con todas esas complicaciones, los usuarios a menudo traducen los resultados en una escala que parece ser más sencilla pero que por lo general carece por completo de sentido: *el porcentaje de cambio en el desempeño*. No son sólo las personas no especializadas quienes lo hacen; he visto hacerlo a académicos reconocidos, aunque nunca a un especialista en medición. Por ejemplo, en la reunión anual de la *American Educational Research Association* (Asociación Estadounidense de Investigación Educativa) realizada en 2006, un importante defensor de la privatización de la educación estadounidense presentó el porcentaje de cambio en los resultados en la Evaluación Nacional del Progreso Educativo para argumentar que en las décadas recientes había mejorado poco el desempeño de los estudiantes. Estaba totalmente equivocado,

como vimos en el capítulo 5, pero para los propósitos actuales, el punto es que su manera de sustentar su conclusión, aunque de sentido común, era disparatada. La razón es que en la mayoría de las escalas el resultado promedio es arbitrario. Considere de nuevo la escala de matemáticas de la prueba SAT. La última vez que se revisó la escala, tenía una media de 500. Suponga que la puntuación de un alumno aumentó de 500 a 600. Eso sería un incremento de 20 por ciento. Pero imagine que al enfrentar una decisión arbitraria, el College Board optó por una media de 400. En ese caso la mejora del estudiante habría sido una ganancia de 25 por ciento, aunque la mejora real del desempeño habría sido idéntica.*

Lo único que desea un usuario común de los resultados de las pruebas es una manera sencilla de describir el logro y se le puede perdonar por verse un poco desconcertado por todas las complejidades aquí descritas. ¿Cómo puede alguien hacer un uso sensato de esas escalas sin tanto alboroto?

La regla fundamental aunque tal vez insatisfactoria es *caveat emptor*: esté al tanto de lo que compra, lo que hacen las escalas de resultados que recibe y lo que no le dicen acerca del desempeño de los estudiantes. Si puede elegir, use la escala que mejor se ajuste a su pregunta. Si quiere saber si un estudiante se mantiene al ritmo de sus pares a medida que avanza por los grados, los equivalentes de grado son una métrica apropiada y práctica. Pero no lo son si lo

* Esto supone que la desviación estándar permanece igual. Técnicamente, el problema es que los resultados de la prueba (como la temperatura Fahrenheit pero a diferencia de la longitud, la velocidad o muchas otras medidas comunes) no son una *escala de razón*, lo cual significa que un cero en la escala de resultados de las pruebas no significa "cero logro". En la mayoría de las escalas, el cero es sólo un punto arbitrario. Incluso en una escala de resultados crudos, en que cero significa que ningún reactivo se respondió correctamente, no necesariamente significa "ningún conocimiento del dominio", sólo quiere decir que no hay dominio del material particular usado en la prueba. El porcentaje de cambio sólo es una métrica con sentido en el caso de las escalas de razón.

que quiere es comparar las mejoras de los niños de secundaria con las mejoras de los estudiantes de primaria.

Además, tenga cuidado de no atribuir a esos reportes más significado del que en realidad tienen o de simplificarlos en algo que parece más sencillo de asimilar. En el caso de los estándares, eso significa que no debe perderse de vista el hecho de que los estándares de desempeño son sólo expresión del juicio de alguien que también podría haberse establecido en un nivel muy diferente, y que las etiquetas que se les asignan pueden venir cargadas de connotaciones excesivas e injustificadas. ■

Capítulo 9
Validez

Los encargados del departamento de admisión utilizan los resultados obtenidos en las pruebas para ayudarse a decidir qué sustentantes tienen mayor probabilidad de éxito en la universidad. Los profesores los utilizan al diagnosticar las fortalezas y debilidades en el aprendizaje de sus alumnos. Al parecer, todos (educadores, padres, reporteros, agentes inmobiliarios) usan los resultados de las pruebas para juzgar el desempeño educativo de las escuelas, estados e incluso países. La ley NCLB requiere de los resultados de las pruebas para determinar qué escuelas ameritan sanciones. Al inicio de este libro sacamos conclusiones acerca del vocabulario de estudiantes universitarios a partir de una hipotética prueba de vocabulario. En cada uno de esos casos, la gente fundamenta una inferencia específica en un determinado resultado de la prueba.

¿En qué medida están esas conclusiones justificadas por los resultados de las pruebas usados para sustentarlas? Como mencioné desde el principio de este libro, esta es la cuestión de la validez, que es el criterio individual más importante para evaluar las pruebas de logro. La importancia de la *validez* se reconoce con la generalidad suficiente para llegar a leyes y regulaciones. Por ejemplo, la ley NCLB exige que las evaluaciones "se utilicen para propósitos para los cuales dichas evaluaciones sean válidas y confiables".

Pero ¿qué significa en realidad "validez"? Parece muy simple; después de todo, el término se emplea en el habla cotidiana. Sin embargo, resulta que hay más en la historia de lo que se aprecia a

simple vista. Muchas de las controversias actuales más importantes acerca de la evaluación, como las disputas respecto de las pruebas de alto impacto, tienen en su raíz desacuerdos referentes a la validez y no pueden ser resueltas sin una consideración más cuidadosa de lo que conlleva la validez.

Este capítulo aclara lo que entendemos por validez, revisa varias de las principales amenazas para una inferencia válida y explica algunos tipos de evidencia que debemos evaluar. Los capítulos siguientes aplican esta noción a tres de las áreas de controversia más importantes en la evaluación actual en Estados Unidos: las pruebas de alto impacto, el sesgo y la evaluación de estudiantes con necesidades especiales.

De manera rutinaria empleamos los términos *válido* y *validez* en todo tipo de contextos. Cuando el ciclista Floyd Landis fue acusado de usar drogas para mejorar el desempeño en su exitoso intento de ganar la Vuelta de Francia, uno de sus abogados afirmó que "la validez de la prueba [de antidopaje] podría ser una de las defensas de Landis".[1] Una columna de opinión publicada en el *Boston Globe* acerca de la controversia que rodeó a la tendencia de George W. Bush de hacer declaraciones al firmar terminaba con el comentario de que "el hecho de que haya estado justificado [en hacer más declaraciones al firmar que sus predecesores] depende de si sus argumentos constitucionales son válidos".[2] En la televisión, en un segmento de la cadena CNN sobre tratamientos para la jaqueca, un experto dijo al respecto de nuevos tratamientos posibles para la migraña que ninguno estaba todavía lo bastante desarrollado para permitir una conclusión válida sobre su eficacia.[3]

Esos tres ejemplos parecen similares, pero entre ellos una distinción sutil adquiere fundamental importancia en la evaluación educativa. El artículo sobre Floyd Landis habla de "validez" para describir a una prueba. Los otros ejemplos emplean la palabra para caracterizar un argumento o una conclusión.

Propiamente, el término *validez* se utiliza en la medición educativa en el segundo sentido, es decir, para describir una inferencia o conclusión específica basada en el resultado de una prueba. La validez no es una característica de la prueba en sí. Esto puede parecer bizantino dado que estamos hablando acerca de conclusiones que se basan en los resultados de la prueba, pero no lo es.

Una razón por la que esta distinción importa es que un determinado resultado de una prueba puede usarse para apoyar una amplia variedad de conclusiones diferentes, algunas de las cuales se justifican y otras no. Presenté un ejemplo en el capítulo 6: las listas de ranking hechas públicas por el Departamento de Educación durante la administración de Reagan. Las listas usaban los resultados promedio de la prueba SAT como medida de la relativa calidad de los programas educativos estatales. Muchos dijimos en el momento que la inferencia acerca de los sistemas educativos estatales era injustificada –no era válida– debido a los muchos otros factores que contribuyen a las diferencias en los resultados promedio de los estados, incluyendo los porcentajes espectacularmente diferentes de alumnos que eligen presentar la prueba. Pero esto no dice nada sobre la validez de la inferencia totalmente distinta que pretendían hacer los diseñadores de la prueba SAT: que los estudiantes que obtienen puntuaciones elevadas en esta prueba tienen mayor probabilidad de un buen desempeño en la universidad. Asimismo, en el capítulo 3 mencioné que E. F. Lindquist, uno de los pioneros más destacados en el desarrollo de las pruebas estandarizadas, argumentó que esas pruebas, de estar bien diseñadas, podían sustentar inferencias útiles acerca de las fortalezas relativas en el desempeño de los estudiantes, pero no sobre la calidad general de un programa escolar. En su opinión, que comparto, la primera inferencia sería más válida que la última.

Una segunda y más polémica razón para considerar que la validez es un atributo de una inferencia y no de la prueba en sí es que

la validez depende del uso particular que se haga de la prueba. Por supuesto, diferentes usos conllevan inferencias distintas. Pero el uso de la prueba también es importante por otros dos motivos.

Algunos usos de las pruebas de hecho socavan la validez. Este es un punto importante de controversia. En particular, cuando se pone a las personas bajo presión suficiente para elevar los resultados en una prueba, algunas de ellas se comportarán de maneras que debilitan la validez. Por ejemplo, algunos maestros recurrirán a ciertos tipos de preparación para la prueba que inflan los resultados o de plano a la trampa descarada. El siguiente capítulo examina este problema que ha adquirido fundamental importancia a medida que se han elevado las presiones de las pruebas de alto impacto.

Por último, diferentes usos de las pruebas pueden tener consecuencias distintas. En parte como respuesta al trabajo de dos de los teóricos más destacados de la validez del siglo pasado –Sam Messick del Servicio de Evaluación Educativa (Educational Testing Service, ETS) y Lee Cronbach de la Universidad de Stanford– en la profesión se ha vuelto común considerar que los *efectos* del programa de evaluación son parte de la validez. Este concepto suele denominarse *validez consecuente*. En mi experiencia, esto ha sido una fuente de confusión interminable entre la gente ajena al campo, lo cual difícilmente puede sorprendernos. Los efectos de un programa de evaluación pueden ser malos incluso si la inferencia que se basa en los resultados es válida, y viceversa. Por ejemplo, varias jurisdicciones, incluyendo las ciudades de Chicago y Nueva York, ahora retienen a los estudiantes en ciertos grados si no logran alcanzar el punto de corte en una sola prueba. Esas políticas han provocado un debate vehemente. Un lado argumenta que la "promoción social" –la promoción al siguiente grado basada en la edad más que en el verdadero aprendizaje– engaña a los estudiantes al permitirles avanzar en la escuela sin dominar el material que necesitan; el otro sostiene que retener a los estudiantes en un

grado hace más daño que bien y eleva la probabilidad de que deserten de la escuela antes de terminarla. Ambos lados pueden tener razón, independientemente de la validez de las inferencias acerca del logro –digamos, competencia en matemáticas y lectura– basadas en los resultados particulares que se usaron para dicho propósito.

Hace algunos años presencié el testimonio de una de las más importantes expertas en medición en una audiencia sostenida en el Capitolio para analizar un sistema propuesto de pruebas nacionales. La experta explicaba que dicho sistema tendría importantes efectos negativos no deliberados, expresando su argumento en términos de la validez consecuente. Uno de los políticos más importantes de la audiencia –un gobernador muy activo en la política educativa nacional que apoyaba la idea de las pruebas nacionales– la interrumpió para decir que no entendía a qué se refería con ese término. La experta trató de explicarlo. Fueron de arriba a abajo, sin conseguir nada, mientras la frustración del gobernador aumentaba visiblemente. Dada la naturaleza de la audiencia, nadie podía hablar en voz alta, por lo que me sorprendí pensando en silencio "Dilo: 'Gobernador, me disculpo por el uso de la jerga. Lo que quiero decir es que este programa podría tener efectos negativos importantes no deliberados'". No lo hizo, más bien insistió en usar la frase "validez consecuente", y el gobernador se rindió a la larga, dijo que sencillamente no entendía su argumento y le pidió que continuara. La experta había perdido uno de sus puntos más importantes.

Por lo tanto, en aras de la simplicidad utilizo "validez" para referirme sólo a una inferencia basada en los resultados. Esto no menosprecia la importancia de los efectos de la evaluación. He dedicado más tiempo de mi carrera a investigar los efectos de la evaluación que la mayoría de quienes insisten en que su impacto forma parte de la validez. Es sólo que resulta más claro usar términos

diferentes para referirse a la validez en el sentido clásico –la calidad de la conclusión– y a los efectos de la evaluación.

Como sugiere la cita de la NCLB con que inicié este capítulo, la validez se presenta a menudo como una dicotomía: una conclusión es válida o no. Por desgracia, la situación por lo general es más oscura que eso. La validez es un continuo, uno de cuyos extremos está anclado por inferencias que sencillamente no se justifican. Sin embargo, en el otro extremo del espectro rara vez tenemos la fortuna suficiente para poder irnos de la mesa habiendo decidido que una inferencia es válida, pura y simple. Más bien, algunas inferencias están mejor sustentadas que otras, pero como la evidencia relacionada con este punto suele ser limitada, tenemos que cubrir nuestras apuestas.

Antes de considerar la evidencia utilizada para evaluar la validez, debemos empezar por preguntarnos qué factores pueden socavar la validez y hacer que nuestras conclusiones resulten injustificadas. Existen muchos, por supuesto, pero caen en tres categorías generales: no medir adecuadamente lo que debería ser medido, medir algo que no debería medirse y emplear una prueba de una forma que disminuye la validez. Dejaré el tercer factor para el siguiente capítulo y aquí me concentraré en los dos primeros.

En la literatura técnica, se utiliza el engorroso pero útil término de *sub-representación del constructo* para referirnos al fracaso en medir lo que queremos medir. Esto se remonta a la idea de que una prueba es una muestra de un dominio. Para medir bien el constructo deseado –vocabulario, competencia en álgebra, cualquier cosa– tenemos que hacer un muestreo adecuado del dominio implicado por ese constructo. Por ejemplo, si Zogby sólo hubiese muestreado a los votantes de 45 años tendríamos un caso de subrepresentación del constructo: los jóvenes votan de manera diferente, por lo que su exclusión de la muestra habría dejado sin medir una parte importante del dominio del constructo –el voto

probable de los jóvenes. El constructo –la probable conducta de toda la población de votantes– habría quedado sub-representado. Como se ilustra en este ejemplo, la sub-representación del constructo es un problema para la validez porque es sistemática: algo importante se deja fuera. Si Zogby hubiese muestreado un grupo representativo pero hubiese encuestado a muy pocos votantes, habría tenido mucho error de medición –un margen de error grande– pero no sub-representación del constructo.

Hasta las décadas de los ochenta y los noventa era raro incluir en los programas estatales de exámenes las evaluaciones directas de trabajos en que los estudiantes escriben ensayos que son calificados. Lo común eran las pruebas de opción múltiple de las habilidades de artes del lenguaje. Muchos críticos sostenían, aunque por lo general sin usar el término, que eso era un caso claro de sub-representación del constructo. Es cierto que algunas habilidades que se requieren para la redacción pueden evaluarse por medio de reactivos de opción múltiple. Pero algunas de las habilidades esenciales implicadas por el constructo de "competencia para la redacción" sólo pueden medirse haciendo escribir a los estudiantes. Como resultado, ahora son comunes las evaluaciones directas de los trabajos escritos.

Lo contrario a la sub-representación del constructo es medir algo no deseado. Esto se conoce con el término aún más feo de *varianza irrelevante para el constructo*. Aquí se hace referencia al desempeño de los examinados: una variación en su desempeño es irrelevante para el constructo deseado. Esta varianza no deseada puede tener muchas fuentes. Las tareas de la prueba pueden requerir habilidades que no tienen que ver con el constructo; pueden demandar información antecedente de la que carecen algunos alumnos; factores irrelevantes pueden influir en los jueces; condiciones administrativas pueden tener en algunos estudiantes un efecto diferente que en otros; etcétera. En cada uno de esos casos,

algunos estudiantes tienen un desempeño mejor o peor debido a factores no relacionados con el constructo que pensamos que medimos.

Como esto sugiere, la varianza irrelevante para el constructo puede encontrarse en todo tipo de pruebas, pero es más fácil de ilustrar con evaluaciones complejas del desempeño. Considere una tarea llamada "Densidad" creada por el Consejo de Directores Estatales de Educación (*Council of Chief State School Officers*) para las evaluaciones de ciencia de quinto a octavo grados. El propósito de la tarea, como su nombre lo implica, es evaluar la comprensión que tienen los alumnos del concepto de densidad. Se les pide primero que hagan lo siguiente en un grupo pequeño:

1. Usar una hoja de aluminio de 15 cm de largo x 15 cm de ancho para crear una embarcación que flote en el acuario.
2. Medir el largo y ancho de la parte inferior de la embarcación. Medir su altura. Registrar esta información en la tabla 1 presentada abajo.
3. Medir la masa de la embarcación.
4. Agregar arandelas a la embarcación, una a la vez, hasta que se hunda. Registrar en la tabla 1 la masa de las arandelas agregadas a la embarcación para hacerla hundirse.
5. Repetir el proceso, haciendo embarcaciones de formas distintas y determinando cuántas arandelas aguantan antes de hundirse.

Luego se les pide que respondan por sí mismos algunas preguntas, por ejemplo:

> Suponga que su compañero de laboratorio fabricó dos barcos con dos pedazos idénticos de una hoja de aluminio. El volumen del barco 1 es 500 cm^3 y el del barco 2 de 400 cm^3. ¿Cuál de ellos podrá soportar más masa antes de hundirse? Explique su respuesta.[4]

Cualesquiera que sean sus otras ventajas y desventajas (durante dos o más décadas se han debatido los pros y contras del uso de este tipo de tareas de desempeño complejo para las evaluaciones a gran escala), es claro que esa tarea ofrece mucha oportunidad para que aparezca la varianza irrelevante para el constructo. Por ejemplo, ¿qué sucede a un estudiante que comprende bien el concepto pero que no sabe hacer barcos? (La rúbrica de calificación reconoce este riesgo de manera explícita al advertir que "los barcos pueden estar tan mal construidos que se hundan o se vuelquen antes de que su densidad alcance 1 g/mL"). En otras palabras, las diferencias en la habilidad de los estudiantes para hacer barquitos de aluminio –una destreza que es del todo irrelevante para el constructo que la tarea pretende medir– ocasionará variaciones en el desempeño del estudiante. De igual manera, ¿qué sucede con un estudiante cuya comprensión es correcta pero que es emparejado con otro que no sigue las instrucciones, es indisciplinado o algo por el estilo?

La varianza irrelevante para el constructo también puede surgir de la interacción entre las características de una prueba y las de los estudiantes que la presentan. Por ejemplo, suponga que una prueba se aplica sólo en forma impresa y con letra pequeña. ¿Qué sucederá con los resultados de los buenos alumnos con discapacidades visuales? Su desempeño se verá reducido por su escasa agudeza visual. Eso introducirá en el desempeño una variación adicional que es irrelevante para el constructo que la prueba pretende medir y que no se habría presentado si ninguno de los estudiantes examinados tuviese poca agudeza visual. O suponga que un examen de matemáticas o de ciencia contiene lenguaje innecesariamente complejo. ¿Qué sucederá con los resultados de quienes no son angloparlantes nativos, quienes tienen buen dominio de las matemáticas o la ciencia pero son despistados por esas irrelevantes dificultades lingüísticas? En los capítulos 11 y 12 se considerarán con mayor detalle esas amenazas a la validez.

No hay una prueba de un dominio complejo que pueda ser perfecta. Es inevitable cierta cantidad de sub-representación del constructo y de varianza irrelevante para el constructo, incluso en el caso de una prueba espléndida. Esta es un motivo por el que la mayoría de las inferencias basadas en los resultados de una prueba no pueden ser *perfectamente válidas*. ¿Cómo puede entonces determinarse qué tan válida es una inferencia?

Muchos tipos de evidencia pueden traerse a colación. En casi todas las discusiones del problema se encuentran cuatro tipos de evidencia: análisis del contenido de la prueba, análisis estadístico del desempeño en la prueba, análisis estadístico de las relaciones entre los resultados obtenidos en la prueba y otras variables y las respuestas de los alumnos que presentaron la prueba. Los datos de confiabilidad, aunque a menudo no se presentan como evidencia relacionada con la validez, también son relevantes. Ninguno de esos tipos de evidencia por sí solo es suficiente para establecer que una conclusión es válida, aunque uno solo de ellos puede ser el beso de la muerte que demuestre que una inferencia *no* es válida. Como explico en el capítulo 10, incluso todos ellos juntos son insuficientes cuando hay consecuencias de alto impacto ligadas a los resultados de las pruebas, pero incluso entonces son el punto de partida apropiado.

En casi todo lo que la mayoría de los lectores pueden encontrar, la confiabilidad y la validez se presentan como cuestiones distintas. Por ejemplo, en los reportes técnicos de los programas de evaluación lo común es encontrar capítulos separados para cada uno. Sin embargo, lo cierto es que guardan una estrecha relación.

La confiabilidad es necesaria pero insuficiente para la validez. O, dicho de manera diferente, podemos tener una medida confiable sin validez, pero no una inferencia válida sin confiabilidad. Recuerde que la confiabilidad es sólo la regularidad de la medición. Volviendo al ejemplo de la báscula de baño, suponga que su báscula es

altamente consistente pero consistentemente equivocada. Si se sube y se baja muchas veces, la variación entre las mediciones repetidas será muy pequeña pero el promedio estará de todos modos muy lejos, digamos, unos siete kilos de más. Esto sería una medida confiable, pero la inferencia acerca de su peso no sería válida. Suponga ahora que la escala no tiene sesgo –el promedio a largo plazo de las mediciones, si se pesara muchas veces en sucesión, sería casi correcto–, pero es muy irregular de una ocasión a otra y a menudo varía hasta siete kilos en cualquier dirección. También en este caso, la inferencia sobre su peso, *si sólo se hubiera pesado una vez*, no sería de mucho valor, a pesar de la falta de sesgo. Es decir, su validez sería baja porque usted llegaría a menudo a una conclusión equivocada. Podría obtener una inferencia válida a partir de esta báscula poco confiable si se pesara muchas veces y sacara un promedio, pero eso es así sólo porque la confiabilidad del promedio sería mucho más alta que la de una sola observación.

¿Qué debería uno hacer entonces con las quejas frecuentes sobre el compromiso entre confiabilidad y validez? Uno las escucha a menudo, por ejemplo, cuando la gente discute acerca de las evaluaciones del desempeño frente a los exámenes de opción múltiple. Los defensores de las primeras sostienen a menudo que una preocupación excesiva por la confiabilidad lleva a la gente a inclinarse por formatos como el de opción múltiple (y otros que permiten a los estudiantes responder a muchos reactivos por hora y que requieren una calificación más simple), pero a costa de una menor validez. Su argumento es que las pruebas más confiables no miden algunas cosas que deberían medir. En otras palabras, reclaman que en la persecución de una mayor confiabilidad se empeora la sub-representación del constructo. Hasta cierto punto tienen razón. En ocasiones sucede que para medir ciertas habilidades es necesario usar formatos menos confiables. Sin embargo, esto es cierto con menos frecuencia de lo que muchos piensan;

muchas veces es posible examinar habilidades complejas con formatos como la opción múltiple y no puede contarse con el desempeño en tareas complejas para medir las habilidades de orden superior que a menudo pretenden cubrir. Pero a veces es verdad, como cuando uno quiere medir las habilidades de redacción, y por el momento vamos a considerar ese caso.

Lo que uno tiene en este caso es un difícil compromiso entre confiabilidad y sub-representación del constructo: al disminuir la segunda también se disminuye inadvertidamente la primera. Una modesta disminución de la confiabilidad puede ser un precio razonable por una mejora considerable en la representación del constructo. Pero, más allá de cierto punto, uno se dispara en el pie: termina con una prueba que tiene el contenido deseado pero que produce resultados tan poco confiables que las inferencias que uno desea están debilitadas. No existe respuesta óptima a este problema; dónde se trace la línea, dónde se encuentre el mejor compromiso, depende de cómo use la prueba. Por ejemplo, si yo estuviese elaborando una prueba que va a tener una relación importante con las decisiones sobre estudiantes individuales, yo querría mantener muy alta la confiabilidad, pero estaría dispuesto a ser más indulgente si el propósito de la prueba fuese más descriptivo y diagnóstico.

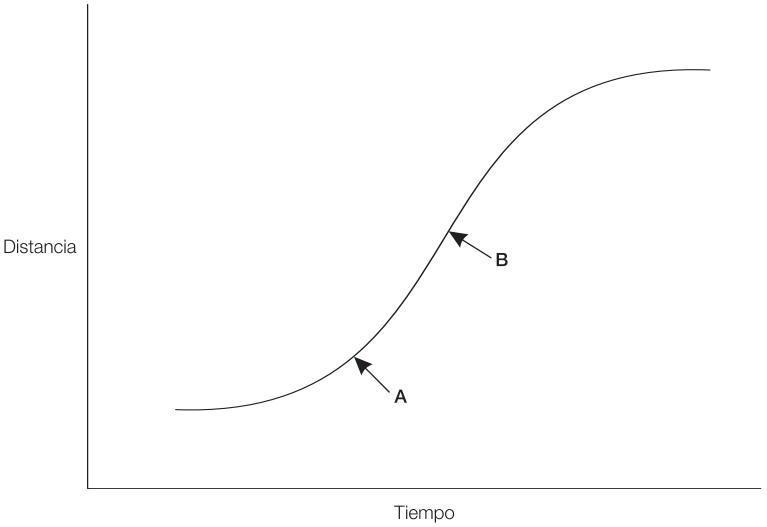
Una segunda conexión entre confiabilidad y validez es que ambas suponen una forma de regularidad. La confiabilidad es la constancia, o posibilidad de *generalización*, del desempeño en casos repetidos de medidas equivalentes, como las veces que usted se sube a su báscula de baño o como en las formas de junio y noviembre de la prueba SAT. Pero la consistencia en medidas *alternativas* del mismo constructo es un elemento clave de la validez. Muchas veces podemos medir la misma cosa de diferentes formas, y no confiamos en esas medidas alternativas si producen hallazgos que difieren de manera considerable. Por ejemplo, los

médicos utilizan dos medidas totalmente distintas para buscar el cáncer de próstata: un examen físico y la prueba en sangre del antígeno prostático específico. Las incongruencias entre los resultados de ambas medidas han sido la fuente de un intenso debate público porque en la medida que sean contradictorias, al menos una está equivocada y se sospecha de la validez de las inferencias acerca de la presencia de cáncer hasta que se entiendan mejor las inconsistencias. Regresaré a esto en el siguiente capítulo porque la regularidad entre medidas es la clave para descubrir la en ocasiones mayúscula inflación de resultados en las pruebas de alto impacto.

Dejando a un lado la confiabilidad, el lugar lógico para empezar a establecer la validez es el contenido de la prueba. Es difícil sostener que una inferencia es válida si se basa en una prueba que incluye el contenido equivocado. Los resultados de un examen casual del contenido de una prueba se conocen como *validez aparente* o validez de fe, como "en apariencia parece válido", pero quienes están en el ámbito de la evaluación no lo consideran como evidencia real de la validez. Más bien, luchan por una evaluación más sistemática del contenido, examinando, por ejemplo, si existen lagunas evidentes en el contenido (sub-representación del constructo), si es apropiado el equilibrio del énfasis, etcétera.

Mucha gente no va más allá del examen del contenido y en la documentación de muchas pruebas se enfatiza la evidencia relacionada con el contenido. Por desgracia, eso no es suficiente para evaluar la validez. Incluso para los expertos suele ser difícil determinar, sólo con ver los reactivos del examen, qué conocimiento y habilidades utilizarán los estudiantes al tratar de responderlos. Detalles aparentemente menores y en ocasiones inadvertidos, como la mala elección de un distractor (una respuesta errónea en un reactivo de opción múltiple) o el uso accidental de un lenguaje innecesariamente complejo, pueden cambiar lo que mide un

Figura 9.1. Reactivo hipotético de matemáticas



reactivo. Además, las habilidades que se necesitan para responder un reactivo dependen a menudo del conocimiento y el entrenamiento previos de los estudiantes. Considere la figura 9.1, que es similar a las figuras que aparecen con frecuencia en las evaluaciones de matemáticas y de ciencia. La gráfica representa el progreso de un individuo durante una carrera a pie, el tiempo se grafica a lo largo del eje horizontal y la distancia cubierta sobre el eje vertical. El reactivo del examen puede pedir al estudiante que identifique que sucede en el punto A (la corredora está acelerando) y en el punto B (mantiene una velocidad constante, la más rápida durante el periodo graficado). Para un estudiante de secundaria que nunca ha visto que una gráfica se use de esta manera, necesita pensar en el problema y traducir velocidad y aceleración en una representación en geometría coordinada. Pero este tipo de representación ahora aparece a menudo en los currículos y algunos estudiantes han visto algo similar muchas veces antes de encontrarlo en la

prueba. Para esos estudiantes, el reactivo mide algo diferente: su recuerdo de algo que aprendieron antes.

La confianza en la validez aparente alcanzó un punto alto durante la oleada de entusiasmo por la evaluación del desempeño, cuando muchos reformadores y educadores supusieron que las tareas complejas necesariamente cubren mejor las habilidades de orden superior que los reactivos de opción múltiple. La gente quería incluir en sus pruebas tareas ricas, realistas e interesantes, por lo que tal vez fue natural que fuesen las tareas (más que otras formas de evidencia que describiré en breve) las que se convirtieran en el requisito de validez para muchas personas que no eran muy conocedoras. En respuesta, Bill Mehrens, que ahora es profesor emérito de la Universidad Estatal de Michigan, acuñó el término "validez de fe", y la investigación confirmó que la confianza en esos tipos de reactivos era en verdad cuestión de fe: el formato de los reactivos no siempre hace una predicción confiable de los tipos de habilidades que los estudiantes utilizarán al abordarlos.[5]

Un segundo acercamiento a la validación es examinar las relaciones entre las puntuaciones obtenidas en la prueba y otras medidas. En algunos casos, podemos evaluar la validez comparando los resultados con un criterio, alguna regla de oro en que podamos confiar. Es probable que encuentre evidencia relacionada con el criterio sobre todo cuando las pruebas se usan para predecir el desempeño posterior. Por ejemplo, la manera convencional de evaluar la validez de las inferencias basadas en los exámenes de admisión a la universidad, como las pruebas ACT y SAT, es ver qué tan bien predicen los resultados de las pruebas el desempeño posterior en la universidad. Lo más común es que el criterio sea el promedio académico en el primer año, pero en algunos estudios es un promedio a más largo plazo, la probabilidad de graduación dentro de un intervalo especificado o alguna otra medida del desempeño en la universidad. Aunque por lo general se denominan

medidas de criterio, la etiqueta es algo engañosa porque las medidas difícilmente están fuera de duda. Calificar, por ejemplo, es muy subjetivo y también presenta marcadas variaciones en la severidad de una disciplina a otra; tiende a ser mucho más estricto en matemáticas y ciencias físicas que en humanidades. Calificar también es vulnerable al sesgo. Una de las maneras estándar de evaluar el posible sesgo en esas pruebas es ver si los estudiantes de un determinado grupo (digamos, las mujeres o los integrantes de grupos minoritarios) obtienen sistemáticamente resultados inferiores o más altos de lo que predicen sus resultados en la prueba. Pero ¿por qué deberíamos asumir, al investigar un sesgo potencial, que la prueba es sospechosa pero no sus calificaciones? Sería igual de razonable suponer que la prueba está menos sesgada y utilizarla para evaluar el sesgo en el proceso de calificar. Pero esas pruebas están diseñadas para sustentar inferencias sobre el desempeño en la universidad, de modo que cualquiera que sea una etiqueta, su capacidad para predecirlo es una base lógica para evaluarlas.

La evidencia relacionada con el criterio es más rara en la evaluación de la educación primaria, secundaria y media superior que en las pruebas de admisión a la universidad, por la sencilla razón de que rara vez tenemos una regla de oro para la comparación. Si la tuviésemos, sencillamente usaríamos esa medida en lugar de la prueba que estamos evaluando. Existen algunas excepciones –por ejemplo, cuando se evalúa si una prueba más corta y menos pesada hace un buen trabajo al replicar los resultados de una más larga y engorrosa. Pero en general, si usted busca la documentación técnica de una prueba que es importante en el distrito escolar de su localidad, verá poca o ninguna referencia a este tipo de evidencia.

Cuando no se dispone de un criterio, la evaluación de la validez se convierte en una tarea más compleja. Un método común es obtener una variedad de medidas diferentes del desempeño, además del que será evaluado. Luego se examinan las relaciones entre

todas ellas, con la esperanza de encontrar que los resultados de la prueba en cuestión tienen una elevada correlación con medidas relacionadas teóricamente y una correlación menos fuerte con medidas con las que uno esperaría que estuviesen menos relacionadas. Es decir, buscamos correlaciones más fuertes con variables con las que debería estar más relacionada si la prueba en cuestión en realidad está midiendo lo que se propone medir. Por ejemplo, los resultados en una nueva prueba de matemáticas deberían tener una mayor correlación con los resultados de otra prueba de matemáticas que con las de una prueba de lectura. Las correlaciones fuertes entre medidas teóricamente relacionadas se conocen como evidencia *convergente* de validez; las correlaciones débiles entre medidas sin relación teórica son evidencia *discriminante*.

La clave es comparar diferentes correlaciones –tanto la evidencia convergente como la discriminante– y no sólo pruebas de contenido relacionado. Esto es de fundamental importancia (lo que también hace difícil evaluar esta forma de evidencia), porque los estudiantes a quienes les va bien en una materia suele irles bien en otra. Como resultado, incluso las calificaciones en materias que consideraríamos no relacionadas suelen mostrar correlaciones altas entre sí. Esto se ilustra en la tabla 9.1, que reporta las

Tabla 9.1. Correlaciones entre resultados de los alumnos de octavo grado en la Prueba de Habilidades Básicas de Iowa

	Lectura	Lenguaje	Habilidades de estudio	Matemáticas
Lectura	1.00	_	_	_
Lenguaje	0.77	1.00	_	_
Habilidades de estudio	0.79	0.80	1.00	_
Matemáticas	0.73	0-75	0.83	1.00

Fuente: A. N. Hieronymous y H. D. Hoover, *Manual for School Administrators, Levels 5-14, ITBS Forms G/H* (Chicago: Riverside Publishing, 1986), Tabla 6.16.

correlaciones entre partes de una antigua edición de la Prueba de Habilidades Básicas de Iowa. Una correlación es una medida de una relación cuyo valor oscila entre -1.0 y +1.0. Un valor de cero significa que no existe relación alguna entre ambas variables. Un valor de 1.0 indica una relación perfecta, lo cual significa que los valores de una variable predicen perfectamente los valores de la otra, y las dos variables son en esencia la misma.* La estatura medida en pulgadas y la estatura medida en centímetros tendrían una correlación perfecta, con un valor de 1.0; proporcionan exactamente la misma información, sólo que en diferentes escalas.

En la tabla, todos los resultados de la Prueba de Habilidades Básicas de Iowa tienen una relación estrecha entre sí. Para dar algún significado a esos valores, considere la correlación de 0.73 entre los resultados en lectura y matemáticas. Esto indica que el sólo hecho de conocer los resultados de los estudiantes en la prueba de lectura le permite predecir alrededor de la mitad de la variabilidad en sus resultados en matemáticas.† Aunque a algunos alumnos les va mejor en una materia que en la otra, muchos de los factores que influyen en su desempeño (salud física, ambiente familiar, algunos aspectos de los antecedentes genéticos, motivación de logro, la calidad de sus escuelas, etcétera) influyen en todas las materias, por lo que los estudiantes que obtienen una alta calificación en una suelen obtener resultados altos también en las otras. Cuando se comparan los resultados promedio de las escuelas, las correlaciones son aún más fuertes. En los mismos datos de

---

* Un valor positivo significa que una variable tiende a incrementarse con la otra; por ejemplo, la estatura tiende a incrementarse con el peso. Un valor negativo indica que una disminuye cuando la otra aumenta; por ejemplo, la velocidad al correr tiende a disminuir con el peso.

† Específicamente, el cuadrado de la correlación ($.73^2 = .53$) es la proporción de la varianza de una variable predicha por la otra. La varianza es una medida específica de variabilidad, el cuadrado de la desviación estándar, que se explica en el capítulo 5.

Tabla 9.2. Correlaciones entre promedios escolares en las pruebas de octavo grado de la Prueba de Habilidades Básicas de Iowa

	Lectura	Lenguaje	Habilidades de estudio	Matemáticas
Lectura	1.00	_	_	_
Lenguaje	0.92	1.00	_	_
Habilidades de estudio	0.94	0.92	1.00	_
Matemáticas	0.88	0.84	0.91	1.00

Fuente: A. N. Hieronymous y H. D. Hoover, *Manual for School Administrators, Levels 5-14, ITBS Forms G/H* (Chicago: Riverside Publishing, 1986), Tabla 6.19.

la Prueba de Habilidades Básicas de Iowa, los resultados promedio de las escuelas en lectura tuvieron una correlación de 0.88 con sus resultados promedio en matemáticas (tabla 9.2), lo que indica que conocer los resultados promedio de las escuelas en una de esas dos materias nos permite predecir más de tres cuartos de la varianza en los promedios de la escuela en la segunda materia. Como resultado, es común encontrar sólo pequeñas diferencias en las correlaciones entre materias relacionadas y no relacionadas, lo que hace difícil el uso de esta evidencia convergente y discriminante.

Esta forma de evidencia fue esencial para valorar la validez en el caso de las evaluaciones de portafolios empleadas en Vermont y Kentucky en la década de los noventa. Ambos estados usaban evaluaciones de portafolios en escritura y matemáticas, pero también aplicaban otras pruebas estandarizadas. En Vermont, las calificaciones del portafolios de matemáticas tenían una correlación casi tan fuerte con una prueba estandarizada de escritura como con una prueba estandarizada de matemáticas.[6] En Kentucky, la evaluación del portafolios de matemáticas tenía una correlación más fuerte con la evaluación del portafolios de escritura que con cualquier otra cosa.[7] Esos hallazgos sugieren que las evaluaciones del portafolios de matemáticas estaban midiendo otras cosas aparte de

las matemáticas –competencia en escritura y diferencias entre maestros en la manera en que se generaban y revisaban las tareas del portafolios.

Dado que muy pocos entienden que los resultados por lo general tienen una fuerte correlación incluso entre diferentes áreas temáticas, las simples correlaciones entre las pruebas en ocasiones se presentan erróneamente como evidencia suficiente de validez. En 2003, un estudio por desgracia muy leído se propuso demostrar que los incrementos en los resultados obtenidos en las pruebas de alto impacto proporcionan una base para hacer inferencias válidas sobre el mejor desempeño de los estudiantes. Los autores escribieron: "El reporte encuentra que el nivel de los resultados obtenidos en las pruebas de alto impacto sigue de cerca al nivel de los resultados obtenidos en otras pruebas, lo cual indica que las pruebas de alto impacto proporcionan información confiable [*sic*] respecto al desempeño del estudiante. Cuando se elevan los resultados obtenidos por un estado en una prueba de alto impacto, deberíamos confiar en que eso representa mejoras reales en el aprendizaje del estudiante."[8] Los autores querían decir "válido" (inferencias justificadas por los resultados) en lugar de "confiable" (resultados consistentes). Entre las evidencias que adujeron para apoyar esta afirmación estaban las correlaciones mostradas en la tabla 9.3. Se trata de correlaciones que oscilan entre 0.35 y 0.96 entre los promedios escolares obtenidos por cada jurisdicción en matemáticas en la prueba de alto impacto y los obtenidos en una prueba de bajo impacto.

Si se miran más de cerca, esas correlaciones proporcionan al menos tantas razones para dudar de la validez de las inferencias basadas en las pruebas de alto impacto como para confiar en ellas. Los autores presentan la evidencia convergente sin la discriminante. El estándar de comparación para esas correlaciones no es cero sino más bien las correlaciones de las pruebas de matemáticas de

Tabla 9.3. Correlaciones entre los resultados promedio en matemáticas obtenidos por las escuelas en pruebas de alto y de bajo impacto, en nueve estados y distritos

Jurisdicción	Correlación
Florida	0.96
Virginia	0.77
Chicago, IL	0.88
Boston, MA	0.75
Toledo, OH	0.79
Blue Valley, KS	0.53
Columbia, MO	0.82
Fairfield, OH	0.49
Fountain Fort Carson, CO	0.35

Fuente: J. P. Greene, M. A. Winters y G. Foster, *Testing High Stakes Tests: Can We Believe the Results of Accountability Tests?* Civic Report 33 (Nueva York: The Manhattan Institute, 2003).

alto impacto en cuestión con otras que en teoría muestran medidas menos relacionadas y que los autores omitieron considerar. La correlación entre las dos pruebas de matemáticas fue .75 en Boston, pero ¿cuál fue la correlación entre la prueba de matemáticas de alto impacto y las pruebas en otras materias? Encontramos un indicio en la tabla 9.2, que también proporciona correlaciones entre promedios escolares, aunque en un contexto de bajo impacto. Con excepción de dos correlaciones, todas las correlaciones que se presentan en la tabla 9.3 entre los resultados de matemáticas en pruebas de alto y de bajo impacto, son menores a las correlaciones que uno encuentra en los datos de la ITBS entre medidas que en teoría *no están relacionadas* –muchas por un margen muy grande. Si una correlación de 0.75 a nivel escolar entre dos pruebas por sí sola es suficiente para establecer validez, se tendría que concluir que la prueba de lectura de la ITBS es una base válida para hacer inferencias sobre matemáticas.

Diversos análisis estadísticos de los datos de las pruebas mismas se usan rutinariamente para evaluar los instrumentos, pero mucho de este trabajo es misterioso y se oculta de la vista de todos, salvo de los usuarios más determinados de los resultados. Por ejemplo, se eleva una bandera roja si los miembros de diferentes grupos –digamos, niños y niñas– que obtienen la misma puntuación en la prueba como un todo muestran un desempeño marcadamente diferente en algunos reactivos. Eso es una señal de posible sesgo en esos reactivos, lo cual podría socavar la validez de las inferencias para uno de los grupos. De igual modo, uno espera que, en promedio, los estudiantes con resultados más altos en la prueba como un todo se desempeñen mejor en cualquier reactivo individual que los que obtuvieron resultados más bajos en la prueba como un todo. Si no es así, el reactivo individual está midiendo algo distinto al resto de la prueba.

Cuando se iniciaron los esfuerzos en la década de los noventa por incluir a más estudiantes con discapacidades en los programas de evaluación de gran escala, varios estudios (incluyendo dos míos) pretendían ver si esas correlaciones "reactivo-prueba" eran similares para los estudiantes con discapacidades y el resto de los alumnos. De no serlo, eso sería una indicación de que la prueba no estaba funcionando bien para los alumnos con discapacidades. (En mis estudios, esas correlaciones estaban bien, pero surgieron otros problemas que pusieron en tela de duda la validez).

El método que se usa con menos frecuencia para evaluar la validez consiste en explorar cómo responden los estudiantes cuando tratan de resolver reactivos individuales. Como mencioné antes, la investigación ha indicado que el formato del reactivo no siempre es un indicador confiable de las habilidades que emplean los estudiantes al tratar de resolverlo. En uno de los mejores estudios de este problema, los investigadores hicieron que los estudiantes explicaran lo que estaban haciendo (por ejemplo, si se

basaban en la aplicación mecánica del conocimiento previo o si empleaban habilidades complejas de solución de problemas) mientras resolvían tareas de opción múltiple y de desempeño en ciencia.[9] Debido a que es demasiado difícil saber qué conocimiento y habilidades usan en realidad los estudiantes cuando responden a los reactivos –y dado que estos pueden variar de un grupo de estudiantes a otro– este tipo de investigación podría hacer una contribución valiosa a la evaluación de la validez. Por desgracia, también es un método costoso, arduo y se lleva mucho tiempo, por lo que es improbable que se convierta en un elemento de rutina de la valoración de los programas de evaluación de gran escala.

¿Qué tanta de esta evidencia relacionada con la validez es probable que encuentre en la información accesible sobre los programas de evaluación con base en pruebas? Si usted está determinado a buscar, casi siempre se presenta la evidencia relacionada con el contenido. Cuánta más pueda encontrar, eso depende de la prueba y de lo profundo que esté dispuesto a excavar. El apetito del público es limitado para, digamos, matrices de correlaciones que muestran evidencia convergente-discriminante, por lo que es poco factible que encuentre una presentación exhaustiva de la evidencia disponible en una ubicación a la mano, como el sitio web del estado o el manual para maestros. Pero incluso si persevera, es poco probable que encuentre todo lo que se describió aquí y a menudo encontrará mucho menos. La pregunta entonces se convierte en otra: ¿La evidencia disponible lo convence de que las conclusiones que le interesan están razonablemente bien sustentadas?

Con todo, esas formas tradicionales de evidencia, no importa qué tan completas sean, no pueden abordar una de las amenazas más formidables para la validez: el riesgo de los resultados inflados, que veremos en el siguiente capítulo. ■

El ABC
de la
evaluación educativa

Capítulo 10
Resultados inflados de las pruebas

Cada año, artículos periodísticos y noticias dadas a conocer por los departamentos de educación del país nos informan que los resultados obtenidos en las pruebas se elevan de nuevo, muchas veces de manera notable. Por lo general, algunos grados o distritos no logran mejoras importantes, y no hay cambio en las brechas en el desempeño entre pobres y ricos y entre mayorías y minorías. No obstante, el argumento principal suele ser positivo: el desempeño está mejorando con rapidez.

Por desgracia, esta buena noticia con frecuencia es más aparente que real. Los resultados obtenidos en las pruebas que se usan para la rendición de cuentas se han inflado, exagerando mucho los verdaderos progresos en el desempeño de los alumnos. Algunas de las mejoras reportadas son por completo ilusorias, otras son reales pero escandalosamente exageradas. Es difícil exagerar la gravedad de este problema. Cuando se inflan los resultados, muchas de las conclusiones más importantes que la gente desprende de ellos resultan equivocadas. Esto afecta a los estudiantes y, en ocasiones, a los maestros.

Este es el sucio secreto de la evaluación de alto impacto. Los periódicos hacen referencias ocasionales a este problema, pero en su mayor parte, los reportes noticiosos y los anuncios de los resultados de las pruebas que hacen los estados y los distritos escolares aceptan sin cuestionar los aumentos en los resultados. El problema también suele ignorarse en el campo profesional de la evaluación.

Si busca una evaluación de la validez de una prueba que le interesa, es probable que encuentre la evidencia con las características discutidas en el capítulo anterior. Esta evidencia es esencial, y en condiciones de bajo impacto puede ser suficiente, pero no le dirá si las mejoras obtenidas bajo presión tienen algún significado. En otras palabras, en el engorroso lenguaje del oficio, no le dirá si las inferencias acerca del mejor aprendizaje basadas en los incrementos de los resultados son válidas.

Hace algún tiempo asistí a una reunión convocada por el personal del *Boston Globe* para analizar el reporte anual del periódico de los resultados de las pruebas en Massachusetts. Alguien planteó el problema de los resultados inflados en las pruebas de alto impacto, sugiriendo que eso podía distorsionar la comparación que hacían los lectores de las escuelas. Dado que yo había investigado este problema desde finales de la década de los ochenta, intervine para tratar de explicar. Un participante que entonces destacaba en los círculos políticos del estado y que ahora es superintendente de un gran distrito escolar de otro estado, me respondió con una frase desdeñosa: "Eso es cuestión de opinión".

Estaba equivocado, no es exclusivamente un asunto de opinión. Si bien sólo en un puñado de jurisdicciones se han realizado estudios creíbles y detallados de la inflación de los resultados, sus hallazgos son bastante sistemáticos y, por lo general, muestran una considerable exageración de las mejoras de los resultados en las pruebas de alto impacto. Varios estudios han comparado con menos detalle los incrementos en las puntuaciones obtenidas en las pruebas estatales y la NAEP. Esos estudios muestran que en muchos casos (pero no en todos) las mejoras en las pruebas del estado, que por lo regular son las que se utilizan para la rendición de cuentas de la ley NCLB, son mucho mayores que las obtenidas en la NAEP.[1] Esas mejoras en pruebas específicas son una señal de inflación. Además, una cantidad importante de estudios ha

documentado conductas de los maestros que pueden ocasionar la inflación de los resultados. En todo el país hay empresas que están vendiendo con entusiasmo materiales que facilitan el trabajo de inflar los resultados, y son muchos los distritos y estados que compran esos materiales para los maestros y alumnos.

Cuando yo y otros que trabajan en este campo señalamos el problema, las reacciones suelen ir de la incredulidad a la ira. De modo que tal vez sea mejor empezar en un terreno menos polémico. También en muchos otros campos puede verse algo similar a la inflación de los resultados de las pruebas (que se conoce también como *corrupción de las medidas* en las ciencias sociales). En efecto, es tan común, que en las ciencias sociales lleva el nombre de ley de Campbell: "Entre más se utilice cualquier indicador social cuantitativo para tomar decisiones sociales, más sujeto estará a las presiones de corrupción y más propenso estará a distorsionar y corromper los procesos sociales que pretende monitorear".[2] De vez en cuando pueden encontrarse en los medios ejemplos de la ley de Campbell que proporcionan un indicio de cómo surge la inflación de los resultados en la evaluación educativa.

Quienes viajan con frecuencia están familiarizados con un ejemplo: las estadísticas de puntualidad de las aerolíneas. Hace varios años, cuando la prensa comenzó a prestar mucha atención a esas estadísticas, muchos viajeros comenzamos a notar que las tasas de puntualidad estaban mejorando, pero no veíamos que llegáramos más pronto a nuestros destinos. En un vuelo largo que tomaba con frecuencia, casi siempre llegábamos "a tiempo", incluso cuando hacíamos largas esperas sobre la pista porque no había puerta disponible para el embarque. El secreto era sencillo: la aerolínea alargó la duración programada del vuelo. Por ejemplo, un artículo publicado en 2000 por el *New York Times* reportó que "[el tiempo programado de] un vuelo del Aeropuerto Internacional Kennedy a Seattle era 22 minutos y 48 segundos más largo que

una década antes, aunque no había cambiado el tiempo en el aire".[3] Una vez que se incrementaron los tiempos programados, las estadísticas de puntualidad mejoraron automáticamente, pero en términos prácticos, la "puntualidad" dejó de significar la misma cosa.

Durante algunos años, el Servicio Postal de Estados Unidos envió correspondencia a una muestra de domicilios para determinar los tiempos de entrega. En principio, esto es igual a la encuesta política y a la prueba de vocabulario del capítulo 2, los tiempos de entrega a la pequeña muestra de domicilios representan los tiempos de entrega en las áreas de las que se extrajeron las muestras. Pero en 1997, las autoridades se enteraron de que los empleados postales de Virginia Occidental habían descubierto las direcciones de la muestra. Para hacer que el servicio postal de su estado pareciera bueno ante sus superiores, los empleados se aseguraron de que esos domicilios recibieran siempre un buen servicio. Para conseguirlo contrataron trabajadores temporales para descubrir las cartas de prueba, de modo que pudiesen apurarse con las direcciones de la muestra. Esto mejoró las estadísticas de tiempo de entrega del estado, pero por supuesto dejó sin cambio los tiempos de entrega de la gran mayoría de los hogares de la región.

El problema de la inflación surge en numerosos campos técnicos en que el funcionamiento de un dispositivo en situaciones complejas reales es simulado por una muestra limitada pero estandarizada. Por ejemplo, a lo largo de los años se ha creado una variedad de pruebas de referencia para evaluar la rapidez de los chips de computadora, y los fabricantes usan los resultados de sus pruebas para la comercialización. Este sistema parece reconfortantemente objetivo, pero tiene dos problemas: diferentes pruebas (diferentes muestras de desempeño) pueden proporcionar resultados distintos y los fabricantes pueden manipular el sistema diseñando chips que se desempeñen bien en las tareas muestreadas por una prueba de referencia. De hecho, algunos fabricantes

fueron acusados de hacer precisamente eso para elevar sus resultados de desempeño a niveles poco realistas, es decir, a niveles que eran superiores al desempeño del chip en una mezcla de tareas reales.[4] En un caso, la acusación fue revertida: el fabricante del chip AMD acusó a un grupo industrial de alterar una prueba de referencia para favorecer a los chips de Intel.[5]

En 1998 surgió una controversia muy similar acerca de las pruebas de emisiones de diésel del gobierno federal. El *New York Times* reportó: "El organismo de Protección Ambiental descubrió que miles de los modernos motores diésel de los camiones de servicio pesado operan limpiamente durante las pruebas de desempeño exigidas por el organismo pero emiten mucha contaminación en el uso normal en carretera... Funcionarios dijeron que... los camiones emiten dos veces más contaminantes de lo que permiten las regulaciones".[6] El problema, una vez más, era si los fabricantes manipularon deliberadamente el sistema, diseñando motores que tuvieran un buen desempeño en la limitada muestra de tareas de la prueba del gobierno a costa de un mal desempeño en el uso real. Un caso similar ocurrido unos años antes que involucraba automóviles llevó a GM a retirar del mercado medio millón de vehículos y al pago de multas.

La corrupción de las medidas de la evaluación también afecta la programación de la televisión. Las "semanas de rastreo" son periodos que ocurren tres veces al año durante los cuales se monitorea la audiencia para contribuir a establecer las tarifas publicitarias. Esas medidas se toman muestreando los hogares y semanas del año y se supone que proporcionan a los anunciantes un cálculo de audiencia que pueden esperar por periodos más prolongados. Para obtener los mejores resultados, los productores dan vida a su programación durante estos periodos de rastreo, añadiendo cualquier material sensacional que puedan para generar un incremento a corto plazo en la audiencia. Si lo logran, la medida será obviamente

engañosa, ya que muchos de los programas perderán espectadores una vez que termine el material inusualmente sensacional.[7]

El ejemplo más perturbador que he encontrado de una medida corrompida fue reportado por el *New York Times* en 2005. La Escuela de Medicina y Odontología de la Universidad de Rochester aplicó una encuesta a cardiólogos del estado. Según el reporte del *Times*: "Una abrumadora mayoría de los cardiólogos de Nueva York dicen que, en ciertas circunstancias, no operan a pacientes que podrían beneficiarse de una cirugía cardiaca porque les preocupa dañar su *ranking* en las tarjetas de calificación médica expedidas por el estado".[8] Por lo menos 83 por ciento de los encuestados dijo que el reporte de las tasas de mortalidad tenía este efecto, y 79 por ciento admitió que "saber que las estadísticas de mortalidad se publicarían" había influido en sus decisiones acerca de realizar la cirugía.*

De modo que no debería sorprender que cuando aumente la presión, los educadores (y para el caso también los estudiantes) se comportarán a veces de maneras que inflan los resultados obtenidos en las pruebas. En realidad, dado lo generalizado del problema en otras áreas, sería sorprendente que ninguno de ellos lo hiciera.

Existen dos maneras distintas de manipular el sistema. Ambas aparecen en los ejemplos anteriores y ambas operan en la evaluación educativa. El primer tipo de manipulación requiere que se distorsione la medida en sí. Hay varias formas de hacerlo. La más sencilla es hacer trampa o mentir acerca de la medición. Para regresar a los

* Esas cifras pueden equivocarse un poco, pero no lo bastante para hacer menos vergonzosos los resultados. Sólo 65 por ciento de los cirujanos muestreados respondieron la encuesta, lo cual es una tasa de respuesta marginalmente aceptable. El riesgo es que los cirujanos que no respondieron pudieran haberlo hecho de manera distinta a los que sí lo hicieron. Pero incluso si la totalidad del 35 por ciento que no respondió hubiese dado una respuesta negativa a esas preguntas (un caso sumamente improbable) aún serían más de la mitad los que dijeron que la publicación de las medidas de mortalidad los llevaron a rehusarse a realizar procedimientos que podrían haber beneficiado a los pacientes.

ejemplos anteriores, yo podría haber generado estimaciones infladas de mejora en nuestra hipotética prueba de vocabulario con sólo cambiar las respuestas de los estudiantes después del hecho, una ventaja a la que, por desgracia, ha recurrido más de un maestro. O en el caso de la encuesta de Zogby, los encuestadores podrían haber sobornado a los votantes muestreados para que dijeran que votarían de una u otra forma, tal vez para generar una ilusión de lo que George H. W. Bush llamó "el gran momento". Pero como veremos, los maestros pueden generar esas distorsiones sin recurrir a la trampa descarada. La manipulación que hicieron las aerolíneas de las estadísticas de puntualidad pueden ser similares: no mintieron acerca de la duración de los vuelos, pero crearon una ilusión de mejora al redefinir lo que, en términos prácticos, significa llegar "a tiempo".

La segunda manera de manipular el sistema es más sutil pero tal vez sea incluso más importante en la evaluación educativa: debilitar la muestra en que se basa la medición, lo cual se ilustra con claridad en la prueba de vocabulario del capítulo 2 y el ejemplo de entrega postal que se presentó antes. En ambos casos, el resultado (el obtenido en la prueba de vocabulario y el estadístico del tiempo de entrega) se calcula usando una pequeña muestra que debe representar algo mucho mayor (todo el vocabulario del estudiante y los tiempos de entrega para todos los domicilios del estado). En el caso de la entrega postal y el ejemplo del vocabulario del capítulo 2, eso dejó de ser cierto cuando abordé a los estudiantes que iban a presentar la prueba y les enseñé las 40 palabras: la medida basada en la muestra ya no representaba al todo mayor.

Esta segunda forma de manipulación es importante no sólo porque es una causa significativa de inflación de los resultados, sino también porque la incapacidad para entenderla es el soporte de las excusas más comunes para la enseñanza inapropiada para la prueba. La gente dirá: "No hay nada malo con los reactivos de la prueba, ¿qué habría entonces de malo en enfocar nuestra enseñanza en

ellos?". Tampoco había nada malo en las direcciones muestreadas por el Servicio Postal Estadounidense en Virginia Occidental o en las 40 palabras de la prueba de vocabulario. El problema no es que el material de la prueba sea malo, sino que ese material es sólo una pequeña muestra de lo que importa. Regresaré a esto en un momento y ofreceré algunos ejemplos concretos de preparación para la prueba que ilustran el problema.

La inflación de los resultados de las pruebas no se convirtió en un problema destacado sino hasta finales de la década de los ochenta, pero hubo indicios anteriores, incluso antes de que la evaluación se convirtiera en la fuente de presión que es en la actualidad. La gente en el campo de la evaluación estaba familiarizada con un patrón "dentado" en las calificaciones o resultados. El primer año que un distrito o estado usaba una prueba nueva, los resultados aumentaban con rapidez. Los incrementos por lo generan disminuían su velocidad después de unos cuantos años. Cuando la jurisdicción reemplazaba la prueba con otra, los resultados mostraban una caída brusca, pero luego repetían el mismo aumento inicial rápido.

Existen dos explicaciones posibles de este patrón. La interpretación caritativa es que los estudiantes amplían sus conocimientos cada vez que se pone en práctica un nuevo examen. De manera gradual dominan una mayor cantidad del material de la prueba antigua, sufren una caída de los resultados por la inclusión del nuevo material en la nueva prueba y luego elevan los resultados al añadir más dominio del nuevo material a su maestría del material del examen anterior. Si esto fuera cierto, no debería preocuparnos la caída en los resultados que ocurre con la introducción del nuevo examen.

La interpretación más escéptica es la sustitución: los estudiantes reemplazan la maestría del material enfatizado por la prueba anterior con el manejo del material enfatizado por la nueva, sin alcanzar en realidad un mayor nivel del dominio mayor del que se extrajo la muestra de la prueba. Uno puede encontrar que una u

otra explicación es más factible, pero el patrón dentado de los datos no indica cuál es la correcta.

Los relativamente pocos estudios que han abordado esta cuestión apoyan la interpretación escéptica: en muchos casos, la maestría del material en la nueva prueba simplemente sustituye a la maestría del material anterior. Una salvedad importante es que, aunque la investigación relevante es congruente, es bastante limitada. Una razón para ello es técnica: se necesita una segunda prueba que sea apropiada para usarse como estándar de comparación con las mejoras en las pruebas de alto impacto, y no siempre se dispone de una. Sin embargo, se enfrenta un obstáculo mayor y es de tipo político: Imagine que usted es el superintendente de las escuelas de un estado que está experimentando grandes mejoras en los resultados obtenidos en las pruebas. Yo u otro investigador nos ponemos en contacto con usted y le solicitamos permiso para explorar la medida en que esas mejoras, que son de enorme importancia política, pueden ser exageradas. Es difícil que se acepte esta investigación y, por tanto, las evaluaciones no se realizan. Muchas veces he argumentado (al parecer con poco efecto) que esta es una cuestión tanto ética como científica: los estudiantes (y para el caso también los maestros) que son sometidos sin consentimiento a programas de evaluación de alto impacto, y el público que paga por ellos y al que se pide que tenga confianza en que el aprendizaje de los estudiantes mejorará; tienen derecho a que se evalúen los efectos de esos programas.

Encontré oposición política a las evaluaciones de los programas de medición de alto impacto desde mi primera incursión en el campo, hace alrededor de dos décadas. A finales de la década de los ochenta, cuatro de nosotros propusimos el primer estudio empírico sistemático del problema en respuesta a una petición de propuestas por parte del departamento de educación de un estado que deseaba valorar su nuevo programa de evaluación de alto impacto.

Después de que presentamos nuestra propuesta, recibí la llamada de una persona del departamento que me explicó que el estado no nos permitiría evaluar la validez de las mejoras en los resultados de las pruebas. Luego obtuve autorización para realizar el estudio en un distrito grande, con la condición de que mantendría al distrito en el anonimato. Estábamos adelantados (las pruebas habían sido ordenadas, las aulas habían sido muestreadas, se había gastado mucho tiempo y dinero) cuando fui convocado a una reunión con el superintendente, quien explicó que el estudio no podía continuar. Dijo que el estudio se publicaría y que si una sola maestra descontenta se daba cuenta de que había sido participante e identificaba el distrito, se vería en problemas con la legislatura de su estado por generar controversia acerca del programa de evaluación.

Más tarde, los cuatro obtuvimos permiso para conducir el estudio en otro lado. El precio de la admisión fue que tuvimos que tomar medidas extraordinarias para proteger el anonimato del distrito, por lo que no puedo decir su nombre, el estado en que se encuentra, ni los nombres de las pruebas que utilizamos. Sin embargo, sí puedo decir que el distrito era grande y que una proporción relativamente alta de sus alumnos eran pobres y pertenecían a grupos minoritarios. También puedo decir que si bien el sistema de evaluación del distrito era considerado de alto impacto según los estándares de finales de la década de los ochenta, para los estándares actuales sería blando. Por ejemplo, no había premios económicos para las escuelas con altos resultados ni amenazas de eliminar a las escuelas o sacar a los estudiantes en respuesta a los bajos resultados o a la incapacidad para mejorarlos. La presión surgía sólo de cosas menos tangibles como la publicidad y la persuasión moral.

Aunque anticuados, los resultados de este primer estudio ameritan una exposición más detallada. Otros estudios muestran inflación de las mejoras en una prueba de alto impacto, pero el nuestro fue el único que verificó si el desempeño en una prueba anterior

disminuía mientras aumentaban los resultados en el nuevo examen; y es posible que, irónicamente, también sea útil justamente porque es anticuado. Uno escuchará con frecuencia que la gente argumenta que un programa de evaluación en realidad no es de "alto impacto" a menos que conlleve recompensas o consecuencias tangibles e importantes para los maestros o los estudiantes. Nuestro estudio (y alguna evidencia adicional que no revisaré aquí) sugiere que este debate es simplemente cuestión de semántica. Si la pregunta es qué se necesita para poner a los maestros bajo presión e inducirlos a inflar los resultados o calificaciones de las pruebas, es claro que no se requieren recompensas o sanciones específicas y tangibles.

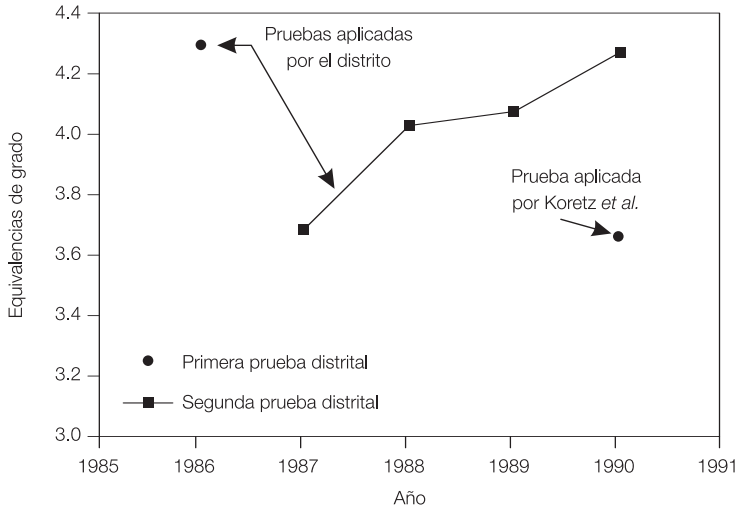
Antes de emprender este estudio, el distrito había experimentado el patrón dentado típico: los resultados habían aumentado por algunos años en una prueba, disminuyeron cuando el distrito adoptó un reemplazo y volvieron a aumentar con rapidez. En la figura 10.1, el diamante en la parte superior izquierda representa los resultados de las pruebas en 1986, el último año que el distrito utilizó lo que llamamos la "primera prueba distrital", que era una de las principales pruebas estandarizadas de logro con normas nacionales. Si la gráfica incluyera años anteriores, usted vería que los resultados en esa prueba habían estado aumentando. En la figura, los cuadrados representan la nueva prueba, llamada "segunda prueba distrital", que empezó a ser utilizada por el distrito en 1987. Puede ver que el desempeño cayó la primera vez que se aplicó la prueba, pero tres años después había regresado al punto elevado anterior.

Una pregunta obvia es ¿qué tan diferentes eran esos dos exámenes? Una vez más, no los puedo identificar, pero puedo decir que se trataba de pruebas tradicionales de logro, estandarizadas, de opción múltiple y que ambas fueron elegidas entre las cinco que en ese entonces dominaban el mercado. Las dos pruebas eran muy similares y se pretendía que representaran un dominio de logro parecido. Sólo diferían en los detalles.

Para entender lo grande que fue el cambio en el desempeño es necesario considerar la escala usada sobre el eje vertical. El desempeño se muestra en términos de equivalencias de grado, una escala descrita en el capítulo 8. En esa escala, el número a la izquierda del punto decimal es el nivel del grado y el número a la derecha es el número del mes en el año académico de 10 meses. Los datos de la figura 10.1 representan el desempeño en matemáticas durante la primavera de cada año, aproximadamente a los siete meses de iniciado el año escolar, por lo que la mediana nacional habría sido una calificación de 3.7 (séptimo mes del tercer grado).

Con esto a la mano, puede ver que en 1986, el último año que el distrito aplicó la primera prueba, el desempeño promedio era un

Figura 10.1. Desempeño en una prueba de impacto moderado y en una prueba de auditoría de matemáticas de tercer grado. Adaptado de D. Koretz *et al.*, "The Effects of High-Stakes Testing: Preliminary Evidence about Generalization across Tests", trabajo presentado en la reunión anual de la Asociación Estadounidense de Investigación Educativa y del Consejo Nacional sobre Medición en Educación, Chicago, abril de 1991



equivalente de grado de alrededor de 4.3, más o menos la mitad de un año académico por arriba del promedio nacional, bastante bueno para un distrito con sus características demográficas. El primer año que se aplicó el segundo examen, los resultados disminuyeron en la mitad de un año académico, lo que hizo disminuir la media del distrito al promedio nacional. Pero los resultados empezaron a subir de nuevo y apenas tres años después el promedio había vuelto al punto en que se encontraba el último año en que se utilizó la primera prueba. De modo que en este caso, el patrón dentado nos deja con una considerable ambigüedad: ¿el desempeño del distrito es similar al promedio o se ubica la mitad completa de un año académico por arriba del promedio?

El nuevo elemento que agregamos –la pieza que ocasionó que se rechazara el primer estudio que propusimos y que se modificara el segundo– fue una prueba adicional. Aplicamos pruebas adicionales a muestras aleatorias de aulas dos semanas después de la evaluación del distrito. Al grupo más grande se le aplicó precisamente la misma prueba que había usado el distrito hasta 1986. Sin embargo, en este caso, a diferencia de 1986, los estudiantes no fueron preparados por sus profesores para esta prueba específica. De hecho, ni los profesores ni los estudiantes sabían que esa era la prueba que se les aplicaría, aunque los maestros sabían que el distrito requería que aplicaran una prueba adicional.*

* Un lector escéptico podría preguntar si no sería posible que los estudiantes tomaran menos en serio esas pruebas complementarias. Ese era un riesgo real. Aplicamos varias pruebas adicionales, una de las cuales era una forma paralela de la segunda prueba distrital, es decir, una forma que empleaba reactivos específicos distintos, pero que estaba diseñada para ser tan parecida a la segunda prueba distrital como fuera factible. Si los estudiantes estuvieran menos motivados para tener un buen desempeño en la prueba complementaria habríamos esperado una caída en el desempeño entre la prueba del distrito y esta forma paralela. Los alumnos de tercer grado (mostrados en la figura) y las niñas de quinto grado pasaron la criba, no así los niños de algunas aulas de quinto grado.

En la muestra a la que se aplicó la primera prueba distrital, que para entonces no había sido usada por el distrito durante cuatro años, los resultados de las pruebas de los estudiantes eran en promedio los mismos que cuando las escuelas aplicaron por primera vez la *segunda* prueba distrital en 1987: por lo menos medio año académico menos que en 1986, el último año que el distrito había usado la primera prueba. Esto es mostrado por el diamante en la parte inferior derecha de la figura 10.1. Por lo tanto, el aumento de los resultados en la segunda prueba no indica que los estudiantes del distrito hayan añadido la maestría del material de esa prueba al manejo del material de la primera prueba. A medida que las cohortes sucesivas de estudiantes mostraban una mejoría progresiva en la segunda prueba, perdían terreno en la primera.

He mostrado la figura 10.1 a cientos, tal vez miles, de personas a lo largo de los años. Muchas veces pregunto, dados esos hallazgos, ¿cuál es el resultado más preciso para proporcionar al público? ¿Están los niños de este distrito medio año académico por arriba del promedio, como indican los resultados del programa evaluativo del distrito una vez que se ha aplicado la prueba por algunos años? ¿O están en el promedio, como sugieren los resultados obtenidos en las pruebas para las cuales no se había preparado específicamente a los alumnos, como sucedió cuando aplicamos la primera prueba en un momento en que ya no se esperaba o con la segunda prueba cuando se aplicó por primera vez en 1987? Muy pocos han tomado el lado de los resultados más altos y una persona, después de escuchar los detalles del diseño del estudio, dijo simplemente: "Eso fue cruel". Pero la abrumadora mayoría, una vez que entendió la idea de que el resultado de una prueba es una muestra de un dominio más grande, ha contestado que las puntuaciones más bajas, las obtenidas en pruebas que no fueron el foco específico de la preparación, hacen una representación más realista del manejo que tienen los estudiantes de todo el dominio.

Y tienen razón, aunque al público siempre se le presenta la estimación más alta. Este estudio fue el primero en demostrar la inflación de las calificaciones o los resultados, análoga a la corrupción de la medición que se hizo en Virginia Occidental de los tiempos de entrega del correo; le siguieron varios estudios que usaban diseños diferentes e investigaban tipos distintos de programas de evaluación; todos mostraron resultados similares.

Tal como advirtió el funcionario del departamento estatal de educación que impidió que se realizara el primer estudio que propusimos, esta investigación saca a flote la cuestión de la validez. Nuestro foco de atención principal era la validez de las inferencias acerca de las mejoras en el aprendizaje de los estudiantes, pero los resultados también pusieron en tela de duda la validez de la inferencia de que el desempeño del distrito estaba muy por arriba del promedio. Las inferencias de este tipo se encuentran entre las más importantes en los sistemas educativos actuales, pero usted no verá mención a una posible inflación de los resultados en las revisiones de la validez de la mayoría de las pruebas, como los reportes técnicos de los programas de evaluación de la mayor parte de los estados. En estos se informará de algunas de las formas tradicionales de evidencia de validez descritas en el capítulo anterior, pero esas no son suficientes para detectar la inflación de los resultados.

Dado el alto impacto del uso actual de las pruebas, sólo podremos confiar en la validez de las inferencias sobre las mejoras si disponemos de un tipo adicional de evidencia relacionada con la validez: una comparación con una segunda medida menos amenazada por la posibilidad de corrupción (que se conoce como prueba de auditoría). La lógica del uso de una *prueba de auditoría* es simple: si las mejoras en la muestra examinada se generalizan al dominio, deberían generalizarse a otras muestras similares del dominio. En este caso, la muestra similar es la prueba de auditoría. También podría ser una muestra diferente de domicilios en

Virginia Occidental o una muestra diferente de votantes en un competidor de la encuesta de Zogby.

Ante nuestros hallazgos, algunos defensores de la evaluación de alto impacto de inmediato culparon al tipo de pruebas usadas en el distrito: ambas estaban compuestas en su totalidad por reactivos de opción múltiple. Por ello era importante proseguir investigando la validez de las mejoras en otros tipos de pruebas. Tuvimos dos oportunidades de hacerlo en Kentucky, en un caso por el interés mostrado por la legislatura del estado y en el segundo por los esfuerzos del subdelegado de educación del estado, Ed Reidy, un hombre de principios que creía que los estudiantes merecen evaluaciones serias de los programas educativos. Kentucky era el lugar ideal para este tipo de investigación, ya que había sido líder en el cambio a las evaluaciones basadas en estándares y utilizaba una diversidad de formatos distintos a la opción múltiple (un formato que abandonó por completo en el curso de algunos años).

El primer estudio en Kentucky examinó lectura de cuarto grado. Kentucky empleaba su propio examen, diseñado por encargo, llamado KIRIS (Kentucky Instructional Results Information System), pero al autorizar el programa KIRIS la legislatura había exigido que el marco de referencia (el documento que especificaba lo que la prueba debía medir) tenía que ser similar al de la NAEP. Dado que se suponía que ambos programas (KIRIS y NAEP) medían aspectos similares de la competencia, la NAEP proporcionaba un estándar evidente de comparación para evaluar las mejoras en KIRIS. Si los estudiantes en verdad estaban aprendiendo más, y no sólo estaban adquiriendo mejores habilidades para presentar esta prueba particular, las mejoras en la prueba KIRIS deberían verse reflejadas de forma importante por las mejoras en la NAEP.

Pero en lectura de cuarto grado, los estudiantes obtuvieron grandes mejoras en la prueba del estado que no recibieron eco en el desempeño de los estudiantes del estado en la prueba de la

Tabla 10.1. Cambio en el desempeño de lectura de cuarto grado en Kentucky en la prueba estatal (KIRIS) y en la NAEP, 1992-1994

	KIRIS	NAEP
Cambios en los resultados escalados	+18.0	-1.0
Cambio en desviaciones estándar	+0.76	-0.03

Fuente: Adaptado de Ronald K. Hambleton *et al.*, *Review of the Measurement Quality of the Kentucky Instructional Results Information System*, 1991-1994 (Frankfort: Office of Education Accountability Kentucky General Assembly, junio de 1995), Tabla 8.1.

NAEP (tabla 10.1). Las dos pruebas fueron reportadas en escalas diferentes, por lo que los resultados no podrían compararse directamente. La forma más sencilla de compararlas era estandarizar ambas pruebas, convirtiendo el cambio a fracciones de una desviación estándar (véase el capítulo 5). La tabla muestra que en dos años los resultados en la prueba KIRIS se incrementaron en alrededor de tres cuartos de una desviación estándar. Esto es un incremento sorprendentemente grande para un tiempo tan corto, lo bastante grande para que quienes estaban familiarizados con ese tipo de datos se dieran cuenta de que algo andaba mal. Sin embargo, el verdadero control fue la comparación con la NAEP: al mismo tiempo que los resultados de la prueba KIRIS aumentaban de manera tan espectacular, los resultados del estado en la prueba de la NAEP mostraron una disminución insignificante.

Este hallazgo no fue obra de la casualidad. En un estudio posterior, un colega y yo investigamos diferentes materias y niveles de grado, y en cada caso encontramos una considerable inflación de los resultados. Algunos casos eran como el de la lectura en cuarto grado: grandes mejoras en la prueba del estado pero ninguna en absoluto en una prueba de auditoría. En otros casos, los estudiantes mostraban alguna mejora en la prueba de auditoría,

pero muy lejana a la observada en la prueba KIRIS. Por ejemplo, en matemáticas de cuarto y octavo grados, las mejoras en las pruebas de la NAEP eran alrededor de un cuarto de las mejoras obtenidas en las pruebas KIRIS, que eran las que se usaban para fines de rendición de cuentas. La situación no parece mejor en la preparatoria, donde las mejoras en los resultados de la prueba KIRIS no se generalizan a las obtenidas en la prueba ACT, que en Kentucky es la prueba de admisión a la universidad que más se utiliza.*

Un estudio parecido exploró el llamado «milagro de Texas», el gran aumento en los resultados en la prueba estatal de alto impacto (TAAS) durante la década de los noventa que se acompañó por una rápida reducción de la brecha entre el desempeño de estudiantes de grupos minoritarios y blancos no hispanos. Dada la penosa persistencia que suele mostrar esta brecha, la reducción de la diferencia en Texas era una noticia importante. Sin embargo, en buena parte era ilusoria. Los estudiantes texanos en verdad mostraron mejoras considerables en la NAEP, pero eran mucho menores que sus mejoras en la prueba estatal TAAS, en algunos casos, la quinta parte. Por su parte, la NAEP no mostró que hubiese una disminución marcada de la brecha entre estudiantes de grupos minoritarios y mayoritarios.[9] Uno esperaría alguna diferencia en las tendencias porque ambas pruebas (TAAS y NAEP) son muy diferentes. No obstante, se traslapan un buen tramo y se pretende que apoyen inferencias algo similares, por lo que las grandes mejoras en la prueba TAAS deberían haberse reflejado de manera más considerable en los resultados obtenidos en la NAEP.

* A diferencia de la prueba SAT, la prueba ACT se diseñó como una prueba de logro y su traslape con la prueba KIRIS es suficiente para convertirla en una prueba de auditoría apropiada, si no es que ideal. Dado que los estudiantes deciden por sí mismos si presentan la prueba ACT, las comparaciones se basaron únicamente en quienes presentaron ambas pruebas.

Por desgracia, es muy poco lo que conocemos acerca de las variaciones en la gravedad de la inflación de resultados. ¿Qué tipos de programas son más susceptibles? ¿Qué tipos de estudiantes o escuelas suelen producir mayor inflación de las calificaciones o resultados? Sospecho que, si todo lo demás permanece igual, es probable que el problema sea peor en el caso de las escuelas que históricamente han obtenido bajos resultados, como muchas de las que atienden sobre todo a estudiantes pobres y de grupos minoritarios. Mi razonamiento es simple: donde los objetivos de desempeño son muy altos en relación con los resultados de las pruebas actuales y donde los apoyos de la comunidad al logro son relativamente débiles, los maestros enfrentan una tarea mucho más difícil y pueden inclinarse más a tomar atajos en su intento por conseguir sus objetivos. Pero la investigación realizada hasta ahora es insuficiente para probar esta hipótesis.

¿Cómo preparan los maestros a sus alumnos para las pruebas en que se les pide rendir cuentas? ¿Cuáles de esos métodos deben considerarse "atajos" en el sentido de que es probable que inflen los resultados?

Durante años, la preparación para la prueba ha sido tema de un intenso debate y se ha utilizado todo tipo de términos diferentes para describir las buenas y las malas formas. Por ejemplo, algunas personas usan "enseñanza de la prueba" para referirse a enseñar reactivos específicos de la prueba (obviamente malo) y "enseñanza para la prueba" para indicar que se enfocan en las habilidades que se supone son representadas por el examen (supuestamente bueno). Sin embargo, otros usan "enseñanza para la prueba" para referirse a la instrucción que se concentra inapropiadamente en los detalles de la prueba (supuestamente malo y con probabilidad de inflar las calificaciones o resultados). Considero que es mejor ignorar todo eso y distinguir más bien entre siete tipos diferentes de preparación para la prueba:

- Trabajar de manera más eficaz
- Enseñar más
- Trabajar más duro
- Reasignación
- Alineación
- Preparación
- Trampas

Los tres primeros son lo que quieren ver los defensores de las pruebas de alto impacto. Es claro que si los educadores encuentran maneras de trabajar con mayor eficacia (por ejemplo, desarrollando mejores programas o métodos de enseñanza) es probable que los estudiantes aprendan más. Hasta cierto punto, si los profesores dedican más tiempo a la enseñanza es probable que aumente el aprovechamiento, y por este motivo la relativa brevedad del año escolar en Estados Unidos nos pone en desventaja en comparación con otras naciones desarrolladas. Sucede lo mismo con la idea de trabajar más duro en la escuela. Por supuesto, eso puede llevarse demasiado lejos. Por ejemplo, no queda claro que privar a los niños pequeños del recreo (cosa que muchas escuelas están haciendo en un esfuerzo por elevar las calificaciones) sea eficaz y, en mi opinión, en cualquier caso es indeseable. De igual modo, si la carga de trabajo de los estudiantes se vuelve excesiva, puede interferir con el aprendizaje y hacerlo motivo de aversión, lo cual podría tener más tarde graves repercusiones. Pero si no se llevan al exceso, cabe esperar que esas tres formas de preparación para la prueba den lugar a mejoras reales en el logro que no sólo aparecerían en los resultados usados para rendir cuentas sino también en otras pruebas y fuera de la escuela.

En el otro extremo, hacer trampa es inequívocamente malo. Ahora son frecuentes los reportes de trampas, las cuales pueden adoptar muchas formas: dar respuestas o indicios a los estudian-

tes durante la aplicación de la prueba, permitirles que cambien sus respuestas después de concluir la prueba, cambiar las respuestas por ellos, brindarles de antemano preguntas de la prueba, etcétera. Las trampas pueden ser o no intencionales, pero independientemente de ello lo único que producen es inflación de los resultados, nunca una mejora real en el aprovechamiento.

¿Qué hay acerca de la reasignación, la alineación y la preparación? Las tres pueden producir mejoras reales, inflación de los resultados o ambas cosas.

La reasignación se refiere a un proceso que ahora es familiar: cambiar los recursos instruccionales (tiempo en el aula, tareas, acoso de los padres, cualquier cosa) para adecuarlos mejor al contenido de una prueba específica. Estudios realizados durante un cuarto de siglo confirman que muchos maestros reasignan la instrucción en respuesta a las pruebas. La reasignación no se limita a los maestros. Por ejemplo, algunos estudios han encontrado que los administradores escolares reasignan a los maestros para ubicar a los más eficientes en los grados en que se aplican pruebas importantes.[10]

¿Es buena o mala la reasignación? ¿Genera mejoras reales en el logro o inflación de los resultados? Depende de dos cosas: a qué resultado se da más énfasis y *cuál se enfatiza menos*. Es claro que cierta cantidad de reasignación es deseable y de hecho es una de las metas de los programas de evaluación. Por ejemplo, si una prueba de matemáticas de noveno grado muestra que los estudiantes tienen un desempeño relativamente malo en la solución de ecuaciones algebraicas básicas, uno desearía que los maestros se esforzaran más en su enseñanza. La cuestión es que esa instrucción es casi un juego de suma cero y dedicar más recursos al tema A implica menos recursos para el tema B.

Los resultados se inflan cuando el tema B (el material que recibe menos énfasis como resultado de la reasignación) también es una parte importante del dominio y, como hemos reiterado, la

mayor parte de las pruebas de logro son una pequeña parte de un gran dominio de logro. Por ende, las pruebas necesariamente omiten material significativo que es importante para la inferencia que los usuarios basarán en los resultados de las pruebas. Eso está bien cuando el impacto es leve: el material incluido representa al omitido, tal como los participantes en la encuesta de Zogby de la que hablamos en el capítulo 2 nos representaron a usted y a mí (y a alrededor de otros 122 millones de individuos), así como a sí mismos.

Pero si los maestros responden a una prueba restando importancia a un material que es importante para esas inferencias pero al que no se da mucho peso en la prueba particular, los resultados de las pruebas se inflarán. El desempeño será más débil cuando los estudiantes presenten otra prueba que enfatice partes diferentes del dominio. Cuando se aplica a los estudiantes una prueba para la que no fueron preparados, una prueba que el distrito había usado hasta unos cuantos años antes, se obtiene el patrón mostrado en la figura 10.1. Esto es una inflación de los resultados del tipo ilustrado por el ejemplo del servicio postal de Virginia Occidental: la muestra examinada deja de representar al dominio.

La alineación es un eje de la política actual de evaluación basada en estándares. Las pruebas deben alinearse con los estándares y la instrucción debe alinearse con ambos. Rara vez se escucha mención de cualquier desventaja de la alineación, y muchos pueden verla como un seguro en contra de la inflación de calificaciones o resultados. Por ejemplo, hace algunos años la directora de una escuela local bien conocida por las altas calificaciones obtenidas por sus alumnos mayoritariamente pobres y de grupos minoritarios, hizo una presentación en la Escuela de Posgrado en Educación de Harvard. En cierto momento, denunció con enojo a los críticos que se preocupaban por la "enseñanza para la prueba". Afirmó que en su escuela no había razón para preocuparse por la enseñanza para la prueba porque el instrumento del estado mide conocimiento y

habilidades importantes. Por lo tanto, si sus profesores enseñan para la prueba, los estudiantes aprenderán cosas importantes. Por supuesto, esta es otra versión del argumento de las "pruebas para las que vale la pena enseñar" que se describió en el capítulo 4.

Esto es una tontería y tuve una corazonada de lo que encontraría si se me permitiera aplicar una prueba alternativa a sus alumnos. La alineación es sólo reasignación con otro nombre. Ciertamente es mejor enfocar la instrucción en material que alguien considera valioso en lugar de derrochar tiempo en cosas triviales. Pero eso no es suficiente. Que la alineación infle los resultados depende también de la importancia del material al que se quita énfasis. La investigación ha demostrado que las pruebas basadas en estándares no son inmunes a este problema; esas pruebas también son muestras limitadas de dominios más grandes, por lo que enfocarse demasiado en el contenido de la prueba específica puede inflar los resultados.

La preparación se refiere a enfocar la instrucción en pequeños detalles de la prueba, muchos de los cuales carecen de significado de peso. Por ejemplo, si resulta que una prueba emplea el formato de opción múltiple para examinar cierto contenido, pueden enseñarse trucos a los estudiantes para trabajar con ese formato; aunque también se les puede enseñar a escribir de maneras adaptadas a las rúbricas de calificación específicas que se usan con una prueba particular. Existe una amplia variedad de métodos de preparación, los cuales también pueden concentrarse en material importante. Por ejemplo, una maestra que enseñaba matemáticas en la secundaria y que participó en uno de mis estudios afirmó que la prueba de su estado siempre usaba polígonos regulares para evaluar el manejo de la geometría plana. Por ende, preguntó, ¿por qué debía molestarse en enseñar sobre los polígonos irregulares? Era un estudio, no una conversación, por lo que no podía darle la respuesta evidente: para que tus estudiantes aprendan algo sobre

ellos. Su foco de atención se había convertido en la prueba, no el currículo más amplio o la inferencia acerca del desempeño que se suponía que la prueba debía sustentar.

La preparación no tiene que inflar los resultados. Si el formato o contenido de una prueba es lo bastante nuevo, una cantidad modesta de preparación puede incluso incrementar la validez de los resultados de la prueba. Por ejemplo, la primera vez que se aplica a alumnos jóvenes una prueba que requiere que rellenen burbujas en una hoja de lector óptico, vale la pena dedicar un tiempo muy breve a familiarizarlos con ese procedimiento antes de empezar el examen.

Sin embargo, es más común que la preparación desperdicie el tiempo o infle los resultados. La inflación ocurre cuando la preparación genera mejoras que se limitan a una prueba específica –o a otras muy similares– que no se generalizan bien a otras pruebas del mismo dominio o al desempeño en la vida real.

Un buen ejemplo es entrenar a los estudiantes para usar un proceso de eliminación cuando responden preguntas de opción múltiple, es decir, a eliminar las respuestas incorrectas en lugar de averiguar la correcta. Un manual *Princeton Review* de preparación para la prueba MCAS de Massachusetts exhorta a los estudiantes a que no lo hagan porque "suele ser más sencillo identificar las respuestas *incorrectas* que encontrar la *correcta*". Luego proporciona un ejemplo inventado que puede ser resuelto por medio de un proceso de eliminación "incluso sin saber nada del tema que pretendía medir el reactivo".[11] Este método no sería posible con un reactivo de opción múltiple bien escrito; de hecho, las preguntas de opción múltiple bien elaboradas usan distractores que atraen a los estudiantes porque reflejan errores comunes y, en tales casos, puede ser difícil eliminar las respuestas incorrectas. Sin embargo, muchos reactivos de opción múltiple no son ideales y esa técnica a menudo ayuda a elevar los resultados.

¿Qué hay de malo en eso? Las mejoras generadas en el desempeño dependen por completo del uso de reactivos de opción múltiple. Todo lo que uno tiene que hacer para que las mejoras desaparezcan es sustituir esos reactivos por reactivos de respuesta construida, los cuales no proporcionan opciones y requieren que los alumnos escriban sus propias respuestas. Por supuesto, cuando los estudiantes necesitan aplicar su conocimiento al mundo real fuera de la escuela, es poco probable que las tareas aparezcan en la forma de un reactivo de opción múltiple. Por consiguiente, preparar a los estudiantes para usar el proceso de eliminación infla los resultados.

Este ejemplo demuestra que la inflación producto de la preparación es en cierto sentido distinta a la que resulta de la reasignación. La reasignación infla los resultados al hacer que el desempeño en la prueba no sea representativo del dominio mayor, pero no distorsiona el desempeño en el material examinado. Los tiempos de entrega en realidad eran más cortos en los domicilios muestreados en Virginia Occidental. En contraste, la preparación puede exagerar el desempeño incluso en el material evaluado. En el ejemplo que acabamos de presentar, los estudiantes a los que se enseña a usar el proceso de eliminación como método para "resolver" ciertos tipos de ecuaciones sabrán menos acerca de esos tipos de ecuaciones de lo que indica su desempeño en la prueba.

La preparación que se enfoca en detalles importantes de una prueba también puede inflar los resultados al crear mejoras en el desempeño que son específicas a la prueba particular. Por ejemplo, considere a la profesora que decidió abandonar los polígonos irregulares. El mundo real no ofrece a los adultos la cortesía de enfrentarlos sólo con polígonos regulares y tampoco lo hacen los autores de otras pruebas. Por ejemplo, la NAEP incluyó polígonos irregulares. De modo que si una prueba sólo utiliza polígonos regulares para evaluar el conocimiento de la geometría plana y si los maestros son lo bastante listos para averiguarlo (por sí mismos

o con la ayuda de los materiales de preparación para la prueba), ellos pueden producir mejoras en el desempeño que se limitan a los polígonos regulares que no se generalizarán a otros exámenes o a las tareas del mundo real que involucren polígonos irregulares.

El ejemplo de los polígonos irregulares puede parecer inverosímil, pero el principio no lo es. Tanto los autores de materiales de preparación para las pruebas como algunos maestros se esfuerzan por identificar los patrones recurrentes en una determinada evaluación, de modo que la preparación resulte factible. Considere el siguiente ejemplo, también de los materiales de *Princeton Review* de preparación para la prueba MCAS: "Siempre que tenga un triángulo recto (un triángulo con un ángulo de 90 grados), puede usar el teorema de Pitágoras. El teorema dice que la suma de los cuadrados de los lados del triángulo (los lados contiguos al ángulo recto) será igual al cuadrado de la hipotenusa (el lado opuesto al ángulo recto)". Eso es seguido por un diagrama de un triángulo recto, con las etiquetas *a*, *b* y *c* en los lados y la ecuación $a^2 + b^2 = c^2$.

Hasta aquí todo bien. Un crítico podría quejarse de que esto es memorista y está fuera de contexto, sin una explicación que dé significado al teorema, pero si el estudiante logra memorizar lo escrito, puede aplicarlo a cualquier triángulo recto en cualquier prueba y a problemas reales cuando salga de la escuela. Pero luego el libro continúa: "Dos de las razones más comunes que se ajustan al teorema de Pitágoras son 3:4:5 y 5:12:13. Como son razones, también funcionarán los múltiplos de esos números, como 6:8:10 y 30:40:50".[12]

Ahora tenemos un problema. Una vez más, el mundo no nos ayuda presentándonos los triángulos rectos cuyos lados tengan largos con la razón 3:4:5 o 5:12:13. Los triángulos rectos pueden aparecer con lados de cualesquier longitud y en cualquier razón, siempre que se conformen a la relación $a^2 + b^2 = c^2$. Nada de esas dos razones es "más común" en el mundo real. Por ejemplo, uno

bien podría encontrar la razón 2:3:3.61. Lo que los autores quieren decir es que son "más comunes en la prueba particular que va a presentar". Los autores de una prueba diferente diseñada para evaluar el mismo estándar podrían usar razones diferentes y, en ese caso, se perdería la mejora en el desempeño lograda por los estudiantes memorizando esas dos razones. Una vez más la inflación de los resultados.

La distinción entre los diversos tipos de preparación para la prueba puede ser confusa. Considere el siguiente ejemplo del condado Montgomery en Maryland (donde mis hijos asistieron a la escuela), como fue advertido en el 2001 por el *Washington Post*: "La pregunta de la hoja de revisión para el examen de álgebra del condado de Montgomery [proporcionada por los funcionarios del distrito] dice en parte: "La cantidad promedio que cada miembro de la banda debe recaudar es una función del número de miembros de la banda, b, con la regla f(b) = 12000/b". En la verdadera prueba, la pregunta dice en parte: "La cantidad promedio que debe pagar cada porrista es una función del número de porristas, n, con la regla f(n) = 420/n".[13] ¿Es un ejemplo extremo de preparación o es una simple trampa? Yo voté por la trampa. Los funcionarios del distrito dieron a los estudiantes una versión apenas disfrazada del verdadero reactivo. Pero incluso si se clasifica mejor como preparación, el resultado final es el mismo: se inflarán los resultados. Si los estudiantes memorizan la solución al reactivo de preparación, pueden "resolver" el reactivo de la verdadera prueba sin ninguna comprensión del álgebra básica, aparte de saber que una incógnita puede ser representada por cualquier letra. Todo lo que el verdadero reactivo les exige es una aritmética ligeramente diferente, la misma operación con distintos números.

Por lo mismo, no siempre queda clara la frontera entre la alineación deseable y la indeseable. Los maestros deberían usar el desempeño en las pruebas para orientar la instrucción, por

ejemplo, enfocándose en el material en que el desempeño de sus alumnos fue relativamente malo. ¿Pero en qué punto cruzan la línea y le roban a Peter para pagarle a Paul?

La prueba del ácido es si las mejoras en los resultados de las pruebas producidas por la preparación para la prueba en verdad representan mejoras significativas en el logro de los estudiantes. No deberíamos preocuparnos por el resultado en una determinada prueba más de lo que nos habríamos preocupado por el conocimiento de los alumnos de las 40 palabras específicas de la hipotética prueba de vocabulario del capítulo 2 o por los votos reales emitidos por los 1,018 votantes encuestados por Zogby el 10 de septiembre de 2004. Deberíamos preocuparnos por la competencia, el conocimiento y las habilidades que se pretende que represente la calificación. Las mejoras que son específicas a una prueba particular y que no se generalizan a otras medidas y al desempeño en el mundo real carecen de todo valor.

Por supuesto, sería posible ignorar simplemente la inflación de los resultados o desecharla como cuestión de "opinión". Eso es precisamente lo que hace la abrumadora mayoría de la gente que hace uso de los resultados de las pruebas. Sin embargo, el costo es grande.

Hacerlo da lugar a una ilusión de progreso y a juicios erróneos acerca del desempeño relativo de las escuelas. Lo más grave: engaña a los estudiantes que merecen una educación mejor y más eficiente. Las alternativas para los educadores, los políticos, los padres y otras personas, son más difíciles. Volveré a ellas en el último capítulo. ■

Capítulo 11
Impacto adverso y sesgo

Cuando yo era un estudiante de posgrado llevé un curso sobre la aplicación e interpretación de las pruebas de inteligencia. Como requisito del curso tenía que practicar la aplicación de dichas pruebas a niños y adultos. Mis sujetos adultos eran mis amigos y sus cónyuges, así como compañeros de mis amigos, todos ellos estudiantes de posgrado o de leyes en universidades sumamente competitivas de Estados Unidos.

Uno de mis voluntarios era un estudiante israelí del posgrado en sociología de una universidad de la Ivy League. Era hijo de un diplomático y, por lo tanto, dedicó mucho tiempo de su niñez al estudio del inglés en escuelas estadounidenses en Europa. Su esposa era estadounidense de nacimiento y estaban criando a sus hijos para que dominaran el inglés y el hebreo. Todo lo cual sirve para decir que su inglés era estupendo, incluso para los elevados estándares de los académicos israelíes.

También era un tipo muy listo, lo cual quedó demostrado por su desempeño en la prueba, a pesar del elevado nivel de ansiedad causado por el hecho de que no sólo estaba presentando la prueba frente a mí (algo de por sí bastante embarazoso) sino también al alcance del oído de su esposa, que también era amiga mía. Sin embargo, sus antecedentes culturales y lingüísticos lo hicieron tropezar varias veces.

Uno de los reactivos preguntaba qué debería hacer el examinado en caso de perderse. El estudiante israelí de inmediato comenzó a decir "Deberías buscar...", luego se puso nervioso y empezó a repetir la frase, sin llegar a completarla. Era claro que la ansiedad dificultaba su búsqueda de la frase que quería en inglés. Le pedí que completara la frase en hebreo y de inmediato dijo que uno debería buscar un "*kever sheikh*". Yo no sabía lo que quería decir, pero una vez que anoté su respuesta pudo poner sus ideas en orden y traducirla para mí: "…la tumba de un jeque".

Eso me paró en seco. Está de más decir que la respuesta no estaba incluida entre las que merecían al menos crédito parcial según el manual. Yo no tenía idea de qué estaba hablando y podría apostar que tampoco la tenían los autores de la prueba. Él me explicó que si uno se pierde en el desierto puede buscar las tumbas de los jeques beduinos para encaminarse, porque sus aberturas se orientan hacia la Meca. En el ambiente en el que él había crecido su respuesta era inteligente y funcional. Pero en la prueba, tal como se suponía que debía ser calificada, su respuesta no merecía crédito porque los autores del instrumento no habían anticipado las respuestas de estudiantes de ese ambiente y es de suponer que desconocían las costumbres funerarias de los beduinos. Y ese no fue el único reactivo que le ocasionó dificultades por el lenguaje o la cultura en que había crecido.

Dadas las circunstancias, transgredí las directrices de la aplicación estandarizada y le otorgué crédito completo por su respuesta acerca de las tumbas beduinas porque era una respuesta razonable en su ambiente natal, de hecho mucho más razonable que algunas de las respuestas estándar que reciben crédito completo. También hice otros ajustes *ad hoc*, como permitirle que presentara en hebreo una parte numérica de la prueba. En este caso, no había inconveniente a mi improvisación, nadie iba a utilizar el resultado para ningún propósito, por lo que yo podía obrar como considerara razonable.

Sin embargo, suponga que yo hubiera aplicado la prueba en las condiciones estándar (sin material presentado en hebreo y sin decisiones *ad hoc* para permitir al examinado que respondiera en ese idioma) y que la hubiera calificado de acuerdo con las rúbricas estándar publicadas. En ese caso, el resultado del estudiante israelí se habría visto disminuido por su desempeño en esos reactivos particulares. La prueba se diseñó para apoyar una inferencia acerca de la inteligencia general del estudiante. Para los propósitos de esa inferencia particular, su resultado habría sido engañosamente bajo, lo que ocasionaría que subestimáramos su inteligencia general.

Este es un ejemplo del *sesgo de la prueba*. Parece obvio en la superficie, pero el concepto de sesgo por lo general se entiende mal, de ahí que valga la pena explicar lo que es y lo que no es el sesgo. Existen tres errores comunes acerca del sesgo en la prueba.

Primero, aunque la gente suele hablar acerca de pruebas sin sesgo (o sesgadas), el sesgo es un atributo de una inferencia específica, no de una prueba. Como expliqué en el capítulo 9, la validez es la medida en que un determinado resultado de la prueba justifica una inferencia particular. El sesgo se refiere a una distorsión sistemática de los resultados que debilita la validez de una determinada inferencia. En el caso del estudiante israelí de posgrado, la distorsión surgió de factores culturales y lingüísticos que disminuyeron su resultado y socavaron la validez de la inferencia acerca de su inteligencia general. El sesgo puede afectar a todo un grupo que presenta una prueba. Por ejemplo, en el capítulo anterior demostré que la inflación de los resultados en una prueba de alto impacto puede sesgar las inferencias acerca de la competencia de la población estudiantil de todo un distrito o un estado. Sin embargo, muchas veces el sesgo afecta sólo a ciertos grupos (como en el caso de mi amigo israelí), produciendo diferencias engañosas en los resultados entre esos y otros grupos.

Una inferencia basada en el resultado obtenido en una determinada prueba puede estar sesgada mientras que otra no. Un ejemplo de esto, que describí en el capítulo 6, son las listas de ranking publicadas por el Departamento de Educación federal al inicio de la administración Reagan. Esas listas usaban los resultados promedio de los estados en la prueba SAT como un indicador de la calidad de sus escuelas. Los datos estaban gravemente sesgados para esta inferencia particular. En algunos estados la prueba fue presentada por la mayoría de los alumnos del último grado de preparatoria mientras que en otros sólo lo hizo un pequeño número. Por ejemplo, en Connecticut 69 por ciento de los alumnos que se graduaban presentaron la prueba SAT en comparación con 7 por ciento en Minnesota. Una razón para esta disparidad fue que en algunos estados la mayoría de las universidades piden resultados de la prueba ACT en lugar de resultados de la prueba SAT, por lo que sólo un pequeño número de estudiantes (sobre todo los que solicitaban ingreso a escuelas muy selectivas fuera del estado) tenían alguna razón para presentar esta última prueba. Por consiguiente, los estados en que menos estudiantes presentaron la prueba habrían obtenido puntuaciones promedio más altas que aquellos en que la presentaban muchos estudiantes, incluso si sus sistemas educativos y poblaciones estudiantiles fueran idénticos, por el solo hecho de que los estudiantes examinados en los primeros estados constituían un grupo más selecto.[1] Sin embargo, el sesgo en esta inferencia en particular no nos dice nada acerca de si hay un sesgo en la inferencia para la cual se diseñó la prueba SAT (la predicción del desempeño de los estudiantes en la universidad). Más adelante presentaré alguna evidencia concerniente a esta última pregunta. (En el capítulo 12 revisaré otro ejemplo de inferencias sesgadas y no sesgadas basadas en un solo resultado: las dos distintas inferencias que pueden sacarse de las puntuaciones del hablante de un segundo idioma en una prueba de admisión a la universidad).

Otra idea errónea frecuente, y tal vez más importante, es que una simple diferencia en los resultados obtenidos por distintos grupos implica sesgo. El resultado del estudiante israelí habría estado sesgado no porque fuera menor de lo que habría sido de haber crecido en Nueva Inglaterra sino por ser *engañosamente* bajo. De igual manera, una diferencia de resultados entre grupos (entre jóvenes pobres y ricos, hombres y mujeres, negros y blancos, asiaticoestadounidenses y blancos) no indica, necesariamente, sesgo. El sesgo puede contribuir o no a la diferencia. Una diferencia en los resultados sólo conlleva sesgo si es engañosa (una vez más, para una inferencia particular).

Voy a emplear de nuevo la prueba SAT como ejemplo. Durante muchos años, las grandes diferencias en los resultados obtenidos en la prueba SAT por diversos grupos sociales han generado un intenso debate. Por ejemplo, los resultados obtenidos en la prueba SAT se incrementan considerablemente con el ingreso familiar (según lo informan los sustentantes) y los resultados promedio muestran una marcada diferencia entre grupos raciales y étnicos. ¿Son sesgo esas diferencias? Más adelante presentaré algunos datos concernientes a las diferencias de grupos raciales y étnicos, pero, por el momento, vamos a considerar de manera hipotética la relación con el ingreso. Es bien sabido que, en promedio, las escuelas que atienden a los niños pobres son de menor calidad que las que atienden a estudiantes de familias de mayor ingreso. Por ejemplo, en las escuelas de las áreas de bajos ingresos los recursos son más limitados y es más probable que los puestos docentes sean ocupados por maestros inexpertos y no certificados. Supongamos ahora (lo que ciertamente no es una suposición riesgosa) que algunas de esas diferencias entre escuelas son importantes y que, como resultado, muchos estudiantes pobres aprenden menos en la escuela y terminan menos preparados para la universidad. Si eso es verdad, las pruebas diseñadas para calcular la preparación

de los estudiantes para la universidad *deberían* arrojar puntuaciones promedio más bajas para los alumnos de esas escuelas de menor calidad y, por lo tanto, para los niños de bajos ingresos. Una inferencia válida acerca de la preparación para la universidad *requiere* que esos alumnos obtengan menores puntuaciones. Qué tan menores deben ser las puntuaciones para proporcionar una predicción precisa es una pregunta abierta. La diferencia que observamos podría ser menor o mayor de lo que debería, lo cual en cualquier caso constituiría sesgo. Por consiguiente, una diferencia de los resultados entre grupos es un motivo para verificar si hay sesgo pero no es una base para asumirlo.

Por supuesto, la posibilidad de sesgo no es la única razón para preocuparse por las diferencias en los resultados de las pruebas entre grupos. Esas discrepancias pueden tener importantes efectos negativos incluso si las inferencias basadas en los resultados están completamente libres de sesgo. Es más probable que los integrantes de los grupos que obtienen bajos resultados no logren alcanzar los estándares exigidos por muchos programas de evaluación de alto impacto en la escuela primaria y secundaria, menos probable que sean admitidos en muchas universidades y todavía menos probable que obtengan empleo en algunos campos.

Es común que a los efectos negativos de las diferencias de grupo en los resultados de las pruebas se les asigne la etiqueta de *impacto adverso*, que es un término más legal que técnico. El impacto adverso puede surgir sin que exista sesgo, y a la inversa: el sesgo puede existir incluso en ausencia de cualquier impacto adverso. Por ejemplo, digamos que los estudiantes de preparatoria encaminados a la universidad que desean convertirse en ingenieros llevan en la preparatoria más cursos de matemáticas que los que quieren especializarse en literatura inglesa, como resultado de lo cual los futuros ingenieros aprenden mucha más matemática. Digamos ahora que en la parte de matemáticas de una prueba de admisión

universitaria ambos grupos obtienen la misma puntuación promedio. En este caso, la ausencia de impacto adverso sería una señal de sesgo, sea porque los resultados fueron engañosamente bajos para los ingenieros o demasiado altos para los especialistas en literatura inglesa. Pero lo más común es que nos preocupe la posibilidad de que un grupo sea perjudicado por la evaluación y en consecuencia nos preocupa la posibilidad de que el impacto adverso, cuando lo encontramos, sea causado o agravado por el sesgo.

El error final acerca del sesgo es una confusión entre sesgo y error de medición. El sesgo y el error de medición son fundamentalmente diferentes y ninguno es causa del otro: un resultado que no es confiable puede no estar sesgado y un resultado sesgado puede ser confiable. Recuerde, la confiabilidad es la consistencia de la medición y el error de medición es la simple falta de consistencia. En el capítulo 7 esbocé una analogía con una báscula de baño barata. Si la báscula no es confiable (si las lecturas incluyen mucho error de medición) entonces habrá mucha inconsistencia en sus lecturas. Pero si la báscula no es confiable pero no tiene sesgo, si se sube suficientes veces y promedia las lecturas, la inconsistencia se eliminará y el promedio se aproximará a la lectura correcta. Sin embargo, si la báscula está sesgada, tenderá a equivocarse en la misma dirección de manera repetida y la lectura promedio a largo plazo será demasiado alta o demasiado baja. Yo tengo un humidificador que es confiable pero está sesgado: de manera sistemática me dice que la humedad relativa de mi recámara es menor de lo que en realidad es. De hecho, dentro de un rango amplio de humedad, su confiabilidad es perfecta: siempre da una lectura de 25 por ciento. El principio general es sencillo: el error de medición se desvanece gradualmente con las medidas repetidas, cosa que no sucede con el sesgo. Cuando en el habla cotidiana decimos que una medida (digamos, una prueba médica) es "exacta", por lo general queremos decir que está relativamente libre de error de medición y de sesgo.

La pérdida de crédito que habría experimentado mi estudiante israelí en la prueba si esta se hubiera aplicado y calificado de la manera convencional, habría sido sesgo más que error de medición. No fue una casualidad, un hecho aleatorio que no habría sucedido si le hubiera aplicado la prueba el miércoles en lugar del martes. Era más bien un problema sistemático: su desempeño fue disminuido por factores lingüísticos y culturales que habrían seguido disminuyendo sus resultados incluso si lo hubiera examinado de nuevo.

El impacto adverso parece más sencillo que el sesgo de la prueba. Calcular el impacto adverso es sólo cuestión de determinar los efectos negativos que sufre un grupo como resultado de las calificaciones de una prueba (como una tasa menor de aceptación por universidades selectivas); y si sólo nos preocupa el impacto adverso, no es necesario molestarse por averiguar si los resultados en verdad son erróneos. Sin embargo, una complicación mal entendida surge en los casos de impacto adverso, y podemos llamarla «el efecto Berkeley».

Berkeley, el campus más antiguo de la Universidad de California, es una de las universidades públicas más selectivas en Estados Unidos. Durante mucho tiempo también ha sido foco de discusiones amargas acerca del impacto adverso y el posible sesgo en las admisiones (que en los años recientes incluye la controversia acerca de la Propuesta 209, la iniciativa estatal que prohibió la acción afirmativa en las instituciones públicas de California y, más recientemente, acusaciones de discriminación en contra de solicitantes asiaticoestadounidenses). Dejando de lado las cuestiones del sesgo y la discriminación, el patrón de impacto adverso en las admisiones (específicamente, la considerable desproporción en la representación de grupos raciales y étnicos entre los grupos de alumnos recién admitidos al primer año) es dramático. En 2006, los afroamericanos constituían menos de 4 por ciento de los estudiantes admitidos (excluyendo el pequeño número que no identificó un grupo racial o étnico). Los latinos conformaban el 14 por ciento;

los blancos 34 por ciento y los asiaticoestadounidenses (que entre esos grupos era el que obtenía las puntuaciones más altas en muchas pruebas estandarizadas) 46 por ciento. Si bien la Propuesta 209 disminuyó considerablemente los porcentajes de afroamericanos y latinos, el patrón de representación desproporcionada era sorprendente incluso en 1997, el último año antes de que se pusiera en efecto la Propuesta 209.[2]

La pregunta evidente es por qué. Algunas personas sugieren que el proceso de admisión, las pruebas de admisión o algún otro aspecto del proceso están particularmente sesgados en contra de los estudiantes de alto logro pertenecientes a grupos minoritarios no asiáticos. No tengo información privilegiada acerca de las admisiones a licenciatura, sea en Berkeley o en otras universidades igualmente selectivas, por lo que no puedo dar una respuesta cabal a esta pregunta. Sin embargo, una parte de la respuesta es clara: este patrón de representación gravemente desproporcionada de grupos con diferentes niveles de desempeño puede surgir sin sesgo alguno. En ausencia de la acción afirmativa o sesgo, la representación desproporcionada se volverá más severa a medida que se incremente la selectividad de la universidad; y recuerde que Berkeley es una universidad muy selectiva.

Hace algunos años asistí a una reunión organizada por la Fundación Ford acerca del impacto adverso en las admisiones universitarias. En un extremo del salón estaba sentado Ward Connerly, el autor de la propuesta 209 así como de otras iniciativas posteriores para prohibir la acción afirmativa en otros estados y en esa época miembro de la Junta de Gobierno de la Universidad de California. En el otro extremo (tanto en sentido literal como figurativo) se sentaron varios abogados con el Fondo para la Defensa Legal NAACP y el Fondo para la Defensa Legal y la Educación de los Mexicoamericanos. Yo me senté en el medio, junto con algunos científicos sociales y otras personas. No es necesario decir que

esta fue una de las reuniones más interesantes a las que he asistido en mucho tiempo, aunque no fue una de las más tranquilas.

Mi tarea era explicar la relación entre la selectividad de las admisiones y el impacto adverso. Mi conclusión era que «el efecto Berkeley» debería ser esperado incluso en la ausencia absoluta de sesgo. A medida que las admisiones se vuelven más selectivas, la representación de los grupos que obtienen bajas puntuaciones (como los afroamericanos y los latinos) va a disminuir de manera progresiva y, exactamente por el mismo mecanismo, aumentará cada vez más la representación de los grupos que suelen obtener altas puntuaciones (como los asiaticoestadounidenses). Esta es una certeza matemática en la medida que la distribución de resultados se ajuste, así sea de manera aproximada, a la curva de campana; específicamente en la medida que los resultados de muchos estudiantes se agrupen cerca del promedio de su propio grupo y que cada vez sean menos las calificaciones que se alejan de ese promedio. Y este efecto es muy poderoso.

Para concretar, presenté al grupo una serie de gráficas basadas en datos simulados que elaboré para imitar una diferencia típica entre estudiantes afroamericanos y blancos. Sólo incluí a esos dos grupos y, para ser realista, hice que el grupo afroamericano fuera más pequeño: 15 por ciento del grupo total de solicitantes. También hice que otros detalles (el tamaño de la diferencia promedio, el tamaño relativo de las desviaciones estándar dentro de cada grupo) fueran realistas, pero no son esenciales y no hay razón para revisarlos aquí. El argumento depende sólo de dos cosas: una diferencia importante en las puntuaciones promedio y una distribución de puntuaciones en que la mayoría de los estudiantes se agrupen cerca del promedio de su grupo.

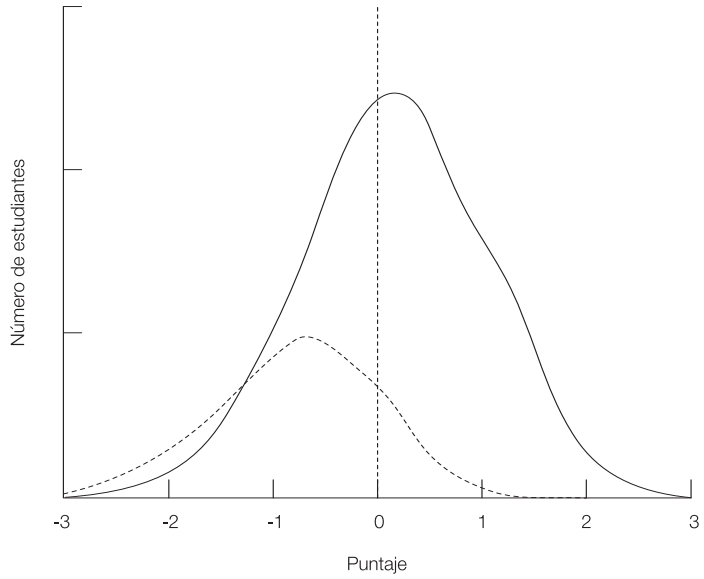
Simplifiqué el proceso de admisión a la universidad y pregunté qué sucedería si las universidades sólo utilizasen un simple punto de corte en una prueba de admisión: si tu calificación se localiza

por debajo del corte, eres rechazado y si se ubica por arriba del corte eres aceptado. No es así como funciona el proceso, o al menos no debería funcionar así; el personal de admisiones debería evitar un punto de corte fijo y buscar numerosos indicadores, no sólo las puntuaciones en una prueba. Pero esta simplificación permite mostrar el problema de manera gráfica y concreta.

Los resultados son notables. Si un sistema no es selectivo –si todos son admitidos sin importar los resultados de la prueba– puede no haber impacto adverso y ambos grupos estarán representados en el grupo de los estudiantes admitidos en proporción a su tamaño en la población: 15 por ciento de afroamericanos y 85 por ciento de blancos. Si se establece un punto de corte en la media global –se admite a todos los estudiantes con calificaciones iguales o mayores al promedio en una prueba, mientras que se rechaza a todos los demás– la representación de los estudiantes afroamericanos disminuye claramente a cerca de 6 por ciento de los estudiantes admitidos. Y un punto de corte en el promedio representa sólo un nivel modesto de selectividad. Para concretar: la puntuación promedio en la parte de matemáticas de la prueba SAT para el grupo que se graduaba en 2006 fue de 518.[3] Usar la media como punto de corte da por resultado una severa sub-representación de los estudiantes afroamericanos, cuya representación en el grupo admitido (6 por ciento) equivale apenas a 40 por ciento de su participación en el grupo de solicitantes (15 por ciento).

Este resultado se muestra en la figura 11.1, donde la línea vertical punteada representa el punto de corte: todos los que se encuentren a la derecha de la línea (con calificaciones por arriba del corte) son admitidos mientras que se rechaza a todos los que se ubiquen a la izquierda de la línea. Las calificaciones se encuentran en una escala de puntajes z (véase el capítulo 5) con una media de cero y una desviación estándar de uno. El punto de corte (la línea vertical) se establece por ende en un puntaje de cero. Puede ver que el

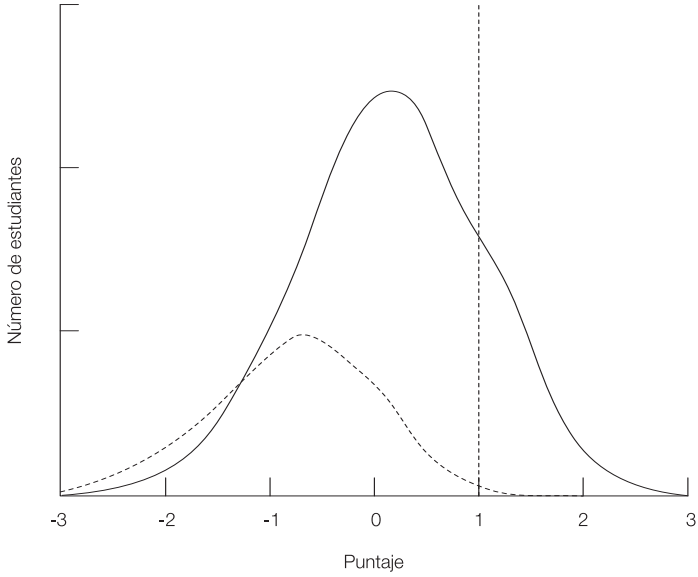
Figura 11.1. Ilustración del impacto adverso con un punto de corte establecido en la media global; los blancos son representados por la línea continua, los negros por la línea punteada



grupo a la derecha de la línea es desproporcionadamente blanco en comparación con la población total de sustentantes.

¿Qué sucede si se hace más selectivo al sistema? Considere elevar el punto de corte a una desviación estándar por arriba de la media, lo que corresponde a una puntuación de 633 en la parte de matemáticas de la prueba SAT. Esto representa un nivel bastante alto de selectividad, aunque no para los estándares de las universidades más selectivas. (Por ejemplo, según la clasificación de universidades del *U.S. News and World Report*, una puntuación combinada de 1380 en las porciones verbal y matemática ubica a un estudiante apenas en el rango percentil 25 entre los estudiantes de primer año recién inscritos en Princeton.)[4] Con este punto de corte, casi todos los estudiantes afroamericanos serían rechazados

Figura 11.2. Ilustración del impacto adverso con un punto de corte establecido en una desviación estándar por arriba de la media global; los blancos están representados por la línea continua, los negros por la línea punteada



y los admitidos sólo constituirían el 1 por ciento del grupo admitido (figura 11.2).

Como muestran las figuras 11.1 y 11.2, la sobre-representación y la sub-representación de los grupos son sólo una función de la forma de la distribución de puntuaciones, en que la mayoría de los estudiantes se congregan cerca del promedio de su grupo. Debido a ello, a medida que se eleva el corte la representación del grupo con bajas puntuaciones en el grupo de estudiantes admitido cae con más rapidez que la representación del grupo con mayores puntuaciones.

Las figuras no muestran las puntuaciones de los asiaticoestadounidenses, pero el incremento de la selectividad tiene en ellos el impacto contrario: incrementa su sobre-representación en el grupo admitido. El mecanismo es el mismo: así como la representación

de los negros cae en relación a la de los blancos porque la puntuación promedio de los primeros es más baja, la representación de los blancos cae en relación a la de los asiáticos porque la puntuación promedio de los blancos es menor que la de los asiáticos.

La lección de este ejercicio es que un impacto adverso serio no necesariamente indica sesgo. Muchos otros factores influyen en las admisiones a Berkeley, como en todas las escuelas selectivas, y la mezcla de estudiantes que admite la universidad refleja algo más que «el efecto Berkeley». Por otra parte, la existencia de dicho efecto no es razón para desechar las preocupaciones por un posible sesgo. En los procesos de admisión de cualquier universidad podría existir sesgo a favor o en contra de cualquier cantidad de grupos –por ejemplo, de los legados (los hijos de los ex-alumnos) y atletas, así como integrantes de los grupos minoritarios. Sin embargo, lo que el ejercicio demuestra es que la severidad del impacto adverso no es suficiente para decirnos si el sesgo está presente.

Dado que ni siquiera el impacto adverso severo indica necesariamente sesgo, ¿cómo puede determinarse si en realidad está presente? En algunos casos, como el de mi amigo israelí, la existencia del sesgo es bastante clara. Por desgracia, en muchos casos no lo es y en ocasiones nos quedamos con la incertidumbre de si una diferencia en el desempeño representa sesgo o una verdadera diferencia en los niveles de competencia.

El primer paso al tratar de identificar el sesgo potencial es examinar el contenido de los reactivos de la prueba, para ver si el contenido o incluso la redacción podrían distorsionar el desempeño de ciertos grupos. Por ejemplo, uno no querría usar el vocabulario de la navegación en problemas escritos de matemáticas si se va a aplicar una prueba a niños pobres de zonas sin salida al mar de la región central del país, porque el hecho de no conocer el significado de *acollador* podría oscurecer la competencia en aritmética. Los dobles negativos son un riesgo si la prueba va a aplicarse a

hablantes no nativos del inglés en cuya lengua materna (por ejemplo, el ruso y el hebreo) un doble negativo (yo no sé nada) es la forma apropiada de expresar una negación (yo sé nada). También sería deseable evitar el lenguaje que algunos estudiantes encuentran ofensivo incluso si lo comprenden, no sólo para evitarles la ofensa sino también para impedir una reacción negativa que pudiera abatir su desempeño.

Revisar el contenido de la prueba en búsqueda de material que pudiera ofender o generar sesgo ahora es una parte rutinaria del desarrollo de las pruebas en la mayor parte de los programas de evaluación de alta calidad. Es necesario, pero no siempre funciona. Así como el análisis del contenido no basta para asegurar que una prueba mide lo que se pretende evaluar, tampoco es suficiente para asegurar que no existe sesgo en los reactivos y puede resultar en el rechazo de reactivos que en realidad no mostrarían sesgo en la práctica.

Por lo tanto, debemos regresar a la evidencia empírica. ¿Qué sucede en realidad cuando se aplica la prueba? ¿Existen patrones de desempeño –generales o para grupos particulares de estudiantes– que sugieran sesgo? El desempeño se examina no sólo para la prueba como un todo, sino también en reactivos individuales.

Una forma común de examinar el desempeño en reactivos individuales recibe el engorroso nombre de *funcionamiento diferencial del reactivo* o DIF (por sus siglas en inglés), el cual se refiere a diferencias de grupo en el desempeño en un reactivo particular entre estudiantes que son comparables en términos de su competencia general. Por ejemplo, considere las diferencias de género. Esas diferencias varían de forma notoria de una prueba y de una muestra de estudiantes a otra, pero más a menudo, las mujeres superan a los varones en las pruebas de vocabulario y de lectura, mientras que los varones tienden a sobrepasar a las mujeres en las pruebas de matemáticas. Esas diferencias no suelen ser muy grandes, pero

aparecen de manera muy sistemática. (Sin embargo, en los años recientes casi ha desaparecido la diferencia a favor de los niños en las pruebas de matemáticas de cuarto y octavo grado de la Evaluación Nacional del Progreso Educativo).[5] Suponga que revisamos el desempeño en una prueba de lectura en que las niñas, en promedio, superan a los niños. Elegimos un reactivo y, oh sorpresa, resulta que el desempeño de las niñas fue mejor en ese reactivo. Eso no nos dice nada, después de todo, las niñas superaron a los niños en la prueba, por lo que era de esperar que su desempeño fuera mejor en un reactivo elegido al azar. Pero suponga ahora que *emparejamos* a niños y niñas con base en sus resultados totales y preguntamos si las niñas y los niños *con el mismo resultado de la prueba* tendrán un desempeño diferente en este reactivo. En condiciones ideales la respuesta sería no; una diferencia sustancial en el desempeño de niñas y niños *emparejados* sería un ejemplo de DIF.

Las pruebas para detectar el DIF ahora son muy comunes, en particular para buscar un desempeño diferencial entre grupos raciales y étnicos y entre hombres y mujeres. Y es habitual encontrar cierto grado de funcionamiento diferencial del reactivo. Pero aunque su identificación es un paso en la dirección correcta, no significa que estamos fuera de peligro. Todavía enfrentamos la tarea de determinar *por qué* los estudiantes emparejados en los dos grupos tuvieron un desempeño diferente en el reactivo. El sesgo podría ser la causa de esa diferencia en el desempeño, pero la causa también podría ser otra cosa, como las diferencias en la instrucción. Por ejemplo, los grupos étnicos no se distribuyen de manera uniforme entre las escuelas: los asiaticoestadounidenses se concentran en algunas regiones, los hispanos en otras, los afroamericanos se agrupan en los centros urbanos y en algunas partes del sur, etcétera. En nuestro sistema educativo descentralizado, la instrucción varía de un lugar a otro, no sólo en los programas y libros de texto formales sino también los patrones de los cursos

tomados y la división del alumnado en función del nivel académico. Por consiguiente, el funcionamiento diferencial del reactivo puede surgir de diferencias en la instrucción experimentadas por el estudiante promedio de diversos grupos étnicos. Lo mismo puede suceder en el caso de las diferencias de clase social en el desempeño. Las diferencias de género son otra cuestión (niños y niñas se distribuyen por igual entre regiones y, en su mayor parte, entre las escuelas), pero en el nivel de preparatoria pueden elegir cursos diferentes y eso también puede acarrear diferencias significativas en el desempeño que aparecen como DIF. Si una escuela desanima a las chicas inteligentes de tomar cursos de matemáticas avanzadas, podríamos esperar que el DIF muestre que el desempeño de las niñas es menos bueno en los reactivos que reflejan el contenido de los cursos avanzados en que están sub-representadas. Esto sería una razón para criticar a la escuela, pero no constituye sesgo en los reactivos; estos estarían en lo correcto al demostrar que las chicas aprendieron menos del material enfatizado en los cursos que no tomaron por haber sido desalentadas.

Pero en algunos casos de DIF, el sesgo parece ser el culpable. Por ejemplo, varios estudios han encontrado que cuando los estudiantes con un dominio limitado del inglés son emparejados con hablantes nativos cuya competencia en matemáticas es similar, el desempeño de quienes tienen un manejo limitado del inglés es menos bueno en reactivos lingüísticamente complejos. Es claro que esto parece ser sesgo: su desempeño en esos reactivos es abatido no por su competencia en matemáticas (que es lo que se supone que mide la prueba) sino por problemas para comprender formas complejas del inglés.

Por lo tanto, en cierta medida el DIF al nivel de reactivos individuales es análogo al impacto adverso al nivel de toda la prueba: es una bandera roja que indica la necesidad de mayor investigación, pero no dice por sí misma que hemos encontrado sesgo.

Para jugar a lo seguro, muchas veces los autores de las pruebas descartan reactivos que muestran cantidades muy grandes de DIF, incluso si no pueden identificar su causa. Pero fuera de eso, ese DIF tiene la ventaja de que nos permite ajustar la mira en reactivos específicos que requieren mayor examen.

A menudo, en un intento por identificar un posible sesgo la gente examina patrones en los resultados de pruebas completas en lugar de revisar el desempeño en reactivos individuales. Esto es menos sencillo de lo que parece al principio. Un error común –que se encuentra en las publicaciones de los científicos sociales así como en las discusiones no especializadas sobre el sesgo y que constituyó el *quid* de una conocida decisión legal acerca de la evaluación para la investigación de antecedentes para el empleo– es suponer que el tamaño de las diferencias entre los grupos es un indicador de sesgo. Es decir, si las pruebas muestran diferencias variables entre los grupos (digamos, entre hombres y mujeres o entre afroamericanos y blancos) la suposición es que es probable que las que muestran diferencias más grandes estén sesgadas. Sin embargo, no existe razón para creer que eso sea cierto. Así como las decisiones sobre el contenido influyen en el *ranking* de las medias del país, también pueden tener impacto en el tamaño de las diferencias entre los grupos. Por ejemplo, hace años descubrí que el tamaño de la brecha entre los estudiantes afroamericanos y blancos en la prueba de matemáticas de la Evaluación Nacional del Progreso Educativo variaba considerablemente entre las cinco áreas de contenido que en esa época comprendía el examen. Por ejemplo, la brecha era mucho mayor en medición que en álgebra. El tamaño de la diferencia podía modificarse mediante el sencillo recurso de cambiar los pesos relativos dados a esas cinco áreas de contenido, de la misma manera en que los cambios en el énfasis en diferentes aspectos de las matemáticas cambian las diferencias entre países en las evaluaciones TIMSS o PISA. Pero a menos que

usted quiera hacer una inferencia muy específica acerca de las matemáticas, por ejemplo, una que requiera comparar un énfasis relativo específico en álgebra con el que se hace en medición, no hay razón para concluir que modificar la mezcla en la Evaluación Nacional del Progreso Educativo podría generar o disminuir el sesgo. Simplemente daría una visión un tanto distinta del nivel de competencia en matemáticas.

Además, la confiabilidad puede empañar las comparaciones de las diferencias de grupo en diferentes pruebas. Un efecto del error de medición es que oculta las diferencias entre grupos. Uno puede pensar que los resultados verdaderos son una "señal" y el error de medición un "ruido" aleatorio. A medida que disminuye la razón señal-ruido, se hace más difícil distinguir una del otro. Por consiguiente, a medida que una prueba se vuelve menos confiable, se hace más difícil ver las relaciones entre los resultados y otros factores. Este problema adopta muchas formas, pero una de ellas es que las diferencias de grupo en los resultados parecerán menores en las pruebas poco confiables que en las confiables. Como dijo en broma un colega mío hace años: "La forma más fácil de disminuir las diferencias de grupo en el desempeño es escribir pruebas malísimas y poco confiables".*

Por lo tanto, para identificar el sesgo necesitamos algún tipo de comparación externa, algo distinto a la prueba en cuestión con lo cual podamos comparar los resultados de las pruebas. Pero no siempre lo tenemos. Explico un caso en el siguiente capítulo: cuando probamos diferentes maneras de evaluar a estudiantes con discapacidades, a menudo carecemos de un criterio externo (cualquier

* Técnicamente, esto es así siempre que el estadístico usado para reportar la diferencia sean los puntajes estandarizados. La mayoría de los estadísticos empleados para reportar brechas de los resultados de las pruebas entre grupos son estandarizados, por ejemplo, diferencias de medias y correlaciones estandarizadas.

otra indicación confiable de cómo debería ser el desempeño de esos alumnos) que nos ayude a determinar qué métodos de evaluación proporcionan los indicadores más válidos (es decir, menos sesgados) de la competencia de esos estudiantes.

Un caso en que disponemos de un estándar externo de comparación son las pruebas de admisión a la universidad. Estas pruebas están diseñadas para apoyar la inferencia principal de que los estudiantes que obtienen mayores resultados en las pruebas tendrán un mejor desempeño, en promedio, en la universidad. Ninguna prueba hace una predicción perfecta del desempeño posterior, por lo que podemos esperar que el desempeño de muchos estudiantes en la universidad sea mejor o peor de lo que predicen sus resultados de las pruebas. Eso no constituye sesgo. La inferencia estaría sesgada si el desempeño de algún grupo en la universidad fuera *sistemáticamente* mejor o peor de lo que predicen sus resultados. Al menos en teoría, disponemos de un criterio para evaluar el sesgo en este caso, podemos recabar datos sobre el desempeño de los estudiantes en la universidad y ver si ciertos grupos se desempeñan sistemáticamente mejor o peor de lo pronosticado por sus resultados en las pruebas.

Necesitamos alguna medida del desempeño de los estudiantes en la universidad que sirva como criterio. Una opción lógica son las calificaciones en los cursos universitarios. La mayor parte de las evaluaciones de la validez de las predicciones basadas en las pruebas de admisión a la universidad examinan qué tan bien los resultados predicen el promedio de calificaciones de los estudiantes durante su primer año de estudios en la universidad. En esas evaluaciones, el sesgo suele calcularse a partir de la sobre y subpredicción de las calificaciones. Por ejemplo, si existe un sesgo en contra de un grupo particular, sus resultados harían una subpredicción de su promedio de calificaciones en el primer año de estudios en la universidad, les iría mejor en la universidad (en promedio) de lo

que predicen sus resultados de las pruebas. El más reciente de esos estudios de validación de la prueba SAT examinó el promedio académico en el primer año de los grupos que ingresaron en 1994 y 1995 a 23 universidades de todo el país que variaban en términos de ubicación, tamaño y selectividad, y proporcionó estimaciones de sub y sobrepredicción por raza, origen étnico y género.[6]

Los resultados no fueron los que esperaría la mayoría de la gente. El estudio demostró un ligero sesgo de la prueba SAT en contra de un grupo: las mujeres blancas. El promedio académico obtenido en el primer año por las mujeres blancas fue ligeramente mayor al pronosticado por sus resultados en la prueba SAT, en alrededor de 0.10 puntos de calificación (en una escala en que A+ es 4.3, A es 4.0 y A- es 3.7, B+ es 3.3, y así sucesivamente). Las predicciones para las mujeres de grupos minoritarios se acercaban mucho al verdadero promedio académico obtenido por esos grupos. Los mayores sesgos (aunque eran todavía muy pequeños) fueron a favor de los hombres de grupos minoritarios. Los hombres afroamericanos y los latinos obtuvieron promedios académicos que eran menores en alrededor de 0.15 puntos de calificación a lo pronosticado por sus resultados en la prueba. Las calificaciones obtenidas por los hombres blancos y los asiaticoestadounidenses eran inferiores a lo anticipado en alrededor de 0.05 de un punto.*

¿Qué deberíamos hacer con los hallazgos de este estudio y otros similares? Existen muchas razones para tomarlos con reservas. En el capítulo 9 hice hincapié en que un criterio debe ser una medida en que se confíe, y existen bases para no tener mucha

* En cierta medida, esos ligeros sesgos pueden surgir del método estadístico empleado más que de las características sustantivas de la prueba SAT. Esos estudios imponen un modelo de predicción de una sola regresión lineal en grupos con calificaciones promedio esencialmente diferentes. Este método tenderá a generar subpredicción para grupos con altas calificaciones y sobrepredicción para grupos con bajas calificaciones.

confianza en las calificaciones universitarias. Para empezar, las calificaciones suelen ser muy poco confiables. Los estudios que correlacionan las puntuaciones con las calificaciones obtenidas en la universidad plantean varios problemas analíticos: carecemos de calificaciones para los estudiantes que fueron rechazados o que fueron aceptados pero decidieron no inscribirse, y el rango de calificaciones entre los estudiantes inscritos a menudo tiene severas restricciones. Tal vez más importante, no tenemos razón para confiar en que las calificaciones obtenidas en la universidad están libres de sesgo. Por ejemplo, digamos que nos preocupa la posibilidad de que las pruebas de admisión estén sesgadas en contra de los estudiantes de grupos minoritarios. ¿Por qué deberíamos suponer que las calificaciones que reciben los estudiantes –que por lo general son asignadas por profesores o asistentes de enseñanza que conocen la raza u origen étnico de los estudiantes– son menos propensas al sesgo que los resultados obtenidos en una prueba que, en su mayor parte, se califica sin esa información?

Otra complicación es que existen muchas razones por las que los estudiantes se desempeñan bien o mal en la universidad, algunas de las cuales no tienen nada que ver con el logro académico y las habilidades de razonamiento que miden las pruebas de admisión. Por ejemplo, algunos estudiantes cuyo origen fue desfavorecido quizá no aprendieron las habilidades de estudio apropiadas para el nivel de exigencia impuesto por los cursos universitarios. Hace algunos años, Uri Triesman, que en ese entonces era profesor de matemáticas en Berkeley, exploró a qué se debía la elevada tasa de reprobación de estudiantes de grupos minoritarios no asiáticos en la clase de cálculo de primer año. Uno de los factores que identificó fue que muy pocos de esos estudiantes se daban cuenta de los beneficios de formar grupos de estudio. Él pudo reducir la tasa de reprobación con diversas iniciativas, una de las cuales fue establecer grupos de estudio para toda su clase.

Una vez más, la solución radica en ser más específico acerca de la inferencia que se basa en los resultados de las pruebas. Una cuestión es si las pruebas de admisión a la universidad, tal como existen en la actualidad, proporcionan indicadores sesgados del probable éxito en la universidad de estudiantes de grupos minoritarios no asiáticos. Este estudio sugiere que la prueba SAT no está sesgada para este propósito. Una pregunta totalmente distinta es si la prueba SAT y otros exámenes de admisión ofrecen una estimación no sesgada del potencial de esos estudiantes para tener éxito si las universidades les brindaran apoyos adicionales de varios tipos, como el que Triesman introdujo hace años en su grupo de Berkeley. El estudio aquí citado no responde esta segunda pregunta ni para los estudiantes de grupos minoritarios ni para cualquier otro.

En general, ¿qué tan común y grave es el problema del sesgo en el sentido en que se emplea el término en medición? ¿Qué tan frecuentes son las distorsiones sistemáticas en las inferencias basadas en los resultados de las pruebas? A lo largo de este capítulo se ha planteado que la gente suele asumir que hay sesgo donde no debería hacerlo. Ni una diferencia grande en los resultados entre grupos ni un grave impacto adverso indican necesariamente sesgo, y los autores cuidadosos de pruebas ahora hacen un uso rutinario de la revisión del contenido y de análisis estadísticos para disminuir el potencial de sesgo. En algunos casos en que mucha gente supone que hay sesgo (por ejemplo, en la predicción del desempeño en la universidad a partir de las pruebas de admisión) la evidencia no lo muestra.

No obstante, nada de esto pretende sugerir que el problema del sesgo (y el potencial para el sesgo) sea menor. Me gustaría concluir con cuatro advertencias.

Primero: el hecho de que grandes diferencias de resultados de las pruebas entre grupos no necesariamente indiquen sesgo no

implica que nunca lo hagan. La respuesta apropiada es tratar a esas diferencias como una razón para verificar si existe sesgo.

Segundo: nuestra información acerca del sesgo suele ser incompleta. El sesgo, igual que la validez, es algo escurridizo. Las técnicas para identificarlo son limitadas y las evaluaciones del sesgo potencial suelen ser imperfectas. Además, el hecho de que no haya sesgo para un grupo (o para una inferencia) no por fuerza implica que no exista sesgo para otro. La evaluación del sesgo potencial, igual que otros aspectos de la validación, es un proceso continuo.

Tercero: si bien en algunos contextos el sesgo puede ser menos generalizado de lo que suponen muchos observadores, en otros casos es más común. El ejemplo más importante de esto se revisó en el capítulo 10: la inflación de los resultados. La mayoría de las personas que utilizan los resultados de las pruebas de alto impacto (educadores, políticos, escritores, padres de familia, agentes inmobiliarios) creen que son indicadores no sesgados de un mejor aprendizaje. La evidencia a la fecha sugiere otra cosa: la investigación no sólo muestra que los resultados pueden estar sesgados sino también que el tamaño del sesgo suele ser enorme. Por ejemplo, el sesgo mostrado en la tabla 10.1, que surgió en el curso de apenas dos años, casi alcanzó el tamaño típico de la brecha total de resultados entre negros y blancos. Se desconoce en buena medida si ese sesgo afecta más a algunos grupos que a otros.

Por último, en algunos contextos el sesgo potencial es particularmente generalizado y difícil de abordar. Tal vez el ejemplo más importante es la evaluación de estudiantes con necesidades especiales, que revisaremos en el siguiente capítulo. ■

Capítulo 12
Evaluación de estudiantes con necesidades especiales

En el campo de la medición pocos temas suscitan emociones tan intensas como la evaluación de estudiantes con necesidades especiales, sea porque presentan discapacidades o porque su competencia del inglés es limitada.* Ambos grupos son grandes y el último está creciendo con rapidez. Desde la década de los noventa las políticas estatales y federales se han propuesto aumentar la participación de esos estudiantes en los programas de evaluación de gran escala y asegurar que, en la medida de lo posible, se les evalúe de la misma manera que a otros alumnos. Las leyes federales imponen a los estados considerables requisitos en lo tocante a la forma de examinar a los estudiantes de ambos grupos.

Esas metas son encomiables, pero las dificultades inherentes a la evaluación apropiada de esos grupos son formidables. En 1997

* Una nota sobre la terminología: para evitar las ofensas, ahora se considera políticamente correcto evitar el antiguo término "dominio limitado del inglés" (DLI) y utilizar mejor "aprendiz del inglés" (AI) para describir a los estudiantes cuya lengua materna no es la inglesa y que no han adquirido total fluidez en ese idioma. A muchos les parece que la expresión "dominio limitado del inglés" es desdeñosa por diversas razones (por ejemplo, porque define a un grupo por un déficit, aquello de lo que los estudiantes carecen o que no pueden hacer), y creen que el problema se resuelve mediante el uso de la expresión "aprendiz del inglés". Sin embargo, para muchas cuestiones de la medición, la expresión "dominio limitado del inglés" describe el hecho pertinente de que algunos estudiantes tienen un nivel de competencia en ese idioma que es lo bastante limitado para interferir con su evaluación apropiada y que, por ende, socava la validez de las conclusiones acerca de su conocimiento y sus habilidades. En contraste, para los propósitos actuales "aprendiz del inglés" es un indicio falso, ya que aquí no nos interesaremos por la cuestión de cuáles de esos estudiantes están aprendiendo el idioma. Por lo tanto, y en aras de la precisión, usaré "dominio limitado del inglés" como un término puramente descriptivo sin que pretenda ser despreciativo.

participé en un panel del Consejo Nacional de Investigación (CNI), una rama de investigación de la Academia Nacional de Ciencias, la Academia Nacional de Ingeniería y del Instituto de Medicina; el panel publicó un reporte donde se analizaba la manera de incorporar a los estudiantes con discapacidades a la reforma educativa basada en estándares. Mientras el panel apoyaba la meta de incluir a esos alumnos en las evaluaciones regulares, planteó que: "La participación significativa de los estudiantes con discapacidades en las evaluaciones a gran escala y el respeto a los derechos legales de dichos individuos en ocasiones requieren que se emprendan medidas que están más allá del conocimiento y la tecnología actuales".[1] A pesar del aumento en la investigación realizada desde ese tiempo, la conclusión mantiene hoy su validez y en cierto grado se aplica también a los estudiantes con un dominio limitado del inglés. Peor aún, al menos por el momento no es posible solventar del todo algunos de los dilemas que enfrentamos al evaluar a esos estudiantes. No podemos esperar que, en el corto plazo, todos ellos sean resueltos mediante más investigación y progresos técnicos; y conforme conocemos la mejor manera de examinar a esos niños algunos aspectos de la política actual parecen alejarse de ese objetivo.

Dado que el relato que hago aquí es desalentador, debo dejar claro que apoyo firmemente los esfuerzos por incluir a los estudiantes con necesidades especiales en el currículo general y en las evaluaciones educativas generales, en la medida que eso sea práctico y sensato. Mi interés por la educación de los estudiantes con discapacidades es muy antiguo: cuando recién egresé de la universidad, trabajé con niños con trastornos emocionales, lo que me llevó a estudiar el posgrado sobre el desarrollo infantil atípico.

También tengo experiencia directa en el hecho de no tener dominio completo de un segundo idioma. Cuando era joven, entre mi periodo como maestro de educación especial y mis estudios de posgrado, viví por un corto tiempo en un kibutz al norte de Israel, al pie

de la montaña donde Saúl fue derrotado por los filisteos –un kibutz que mientras escribía este capítulo fue alcanzado por un misil lanzado por Hezbollah. Me esforcé mucho por aprender el hebreo y de hecho adquirí fluidez en ese idioma, al grado de que muchas veces pensaba y soñaba en hebreo al realizar las actividades cotidianas.

Sin embargo, mi capacidad tenía serios límites; para usar una analogía, yo tenía una competencia limitada del hebreo. Cuando cambiaba al hebreo me volvía aburrido, mi vocabulario era demasiado limitado para exponer ideas más interesantes o complejas o incluso para entender a los demás cuando ellos lo hacían. En esos días, los autobuses tenían altavoces que difundían la Voz de Israel; cuando llegaban las noticias, el conductor aumentaba el volumen y los pasajeros se quedaban callados. Cuando el comentarista empezaba a comentar algo que Henry Kissinger había dicho, yo entendía que estaba analizando lo que había dicho Kissinger pero no podía entender lo que este había dicho. Por otro lado, mi sentido del humor era casi inexistente, ya que el humor requiere jugar con las palabras. Todavía hoy, décadas después, recuerdo el primer chiste que entendí en hebreo, era un chiste simple y malo, pero aun así no lo entendí la primera vez. Por supuesto, también era propenso a cometer errores embarazosos, como la ocasión en que sin darme cuenta traté de sobornar a un guardia Uzi-toting porque confundí los verbos para "fotografía" y "pagar", que tienen un sonido parecido. O la ocasión, algunos años después, en que por la misma razón le explicaba al incrédulo primo de mi esposa que yo era el director de una orquesta sinfónica del Congreso de Estados Unidos.

Debemos continuar con la inclusión de los estudiantes con necesidades especiales en los programas de evaluación de gran escala, pero debemos hacerlo con los ojos abiertos, conociendo las brechas en nuestra comprensión y las deficiencias en nuestros métodos para examinarlos.

Estudiantes con discapacidades

No es fácil decidir quién debe considerarse como una persona discapacitada; por ejemplo, ¿en qué punto un problema de visión, una complicación ortopédica, una dificultad para concentrarse o una alteración emocional dejan de ser una molestia para convertirse en una discapacidad? Una encuesta reciente sobre la prevalencia de enfermedades mentales (*National Comorbidity Survey Replication*) desató un intenso debate público y académico en relación con la línea divisoria entre los estados normales y las enfermedades mentales; por ejemplo, entre la desdicha grave y la depresión clínica.[2] El mismo problema surge con otras discapacidades; pero cuando se trata de examinar a estudiantes con discapacidades, el problema va mucho más allá de los desacuerdos sobre la gravedad requerida para el diagnóstico inicial.

En la mayor parte de lo que se ha escrito acerca de la educación de estudiantes con discapacidades, la información sobre su prevalencia refleja un criterio legal: los requisitos para recibir los servicios según la parte B de la Ley para la Educación de Individuos con Discapacidades (*Individual with Disabilities Education Act*, IDEA).* Es más común que la gente emplee el término *identificación* para referirse a la determinación de que, para los propósitos

* IDEA es la ley federal más importante en lo que concierne a la educación de los discapacitados, la personificación de la revolucionaria Ley para Todos los Niños Discapacitados de 1975, conocida por mucha gente por su designación como Ley Pública, P.L. 94-142, fue el primer estatuto federal que estableció el derecho de los estudiantes con discapacidades a recibir una educación pública adecuada. Fue el origen, por ejemplo, del requisito familiar de que esos estudiantes sean educados en el ambiente práctico menos restrictivo. La Parte B de la ley es el programa primario de subsidio que financia los servicios para los estudiantes con discapacidades y el foco de atención principal de los frecuentes y muy promocionados debates acerca de lo adecuado del financiamiento federal para la educación de los discapacitados.

legales, un estudiante tiene una discapacidad y utiliza *clasificación* para referirse a la etiqueta asignada a la discapacidad del estudiante. En años recientes, alrededor de 11 por ciento de los alumnos de 6 a 17 años en todo el país han sido atendidos en función de la Parte B. Sin embargo, este porcentaje varía de manera sorprendente de un estado a otro. En el año escolar 1999-2000, la prevalencia más baja se reportó en Colorado (alrededor de 9 por ciento) y la más alta (casi 16 por ciento) en Rhode Island (véase la tabla 12.1). Los otros estados se distribuyeron a lo largo del continuo entre esos dos estados.

Las diferencias en las tasas de prevalencia reportadas por los estados se vuelven mucho más grandes, y cuando se examinan las categorías individuales de discapacidad, simplemente no son creíbles. La variación en la prevalencia reportada de las discapacidades

Tabla 12.1. Porcentaje de estudiantes de 6 a 17 años atendidos bajo la Parte B de IDEA, estados y Estados Unidos (50 estados y D.C.), categorías seleccionadas de discapacidad, 1999-2000

	Estado con el porcentaje más bajo	Estado con el el porcentaje más alto	Total en EE.UU.
Todas las discapacidades	9.1 (CO)	15.6 (RI)	11.3
Discapacidad específica de aprendizaje	3.0 (KY)	9.1 (RI)	5.7
Habla/lenguaje	1.0 (IA)	3.9 (WV)	2.3
Retardo mental	0.3 (NJ)	3.0 (WV)	1.1
Trastorno emocional	0.1 (AR)	1.9 (VT)	0.9
Discapacidades visuales	0.02 (IA, NJ)	0.08 (TN, UT)	0.05

Fuente: *To Assure the Free Appropriate Public Education of All Children with Disabilities, 23rd Annual Report to Congress on the Implementation of the Individuals with Disabilities Education Act* (Washington, DC: US Department of Education, 2001), Tabla AA10 (*http://www.ed.gov/about/reports/annual/osep/2001/appendix.a-pt1.pdf*. (revisado el 20 de julio de 2006).

específicas de aprendizaje se triplicó de 3 por ciento en Kentucky a 9 por ciento en Rhode Island. La diferencia entre la mayor y la menor prevalencia reportada de retardo mental fue un factor de 10.*

En su mayor parte, esas impresionantes variaciones entre estados no reflejan diferencias reales en la prevalencia de las discapacidades. Más bien, surgen sobre todo de las diferencias en las políticas estatales y locales concernientes a la identificación y clasificación. La investigación ha demostrado que las prácticas de los educadores aumentan todavía más la incongruencia, por lo que incluso bajo las restricciones de las políticas específicas de un estado, es común encontrar variaciones considerables de un lugar a otro que, en su mayor parte, parecen carecer de justificación. Por ejemplo, durante varios años participé en un comité que aconsejaba al Departamento de Educación del Estado de Nueva York acerca de la educación de estudiantes con discapacidades. Un tema frecuente eran las tasas de prevalencia anómalamente altas y al parecer injustificadas que reportaban algunos distritos suburbanos del sur del estado, tasas que eran mucho más altas que cualquiera de la tabla 12.1.

Es evidente la importancia de las incongruencias en las tasas de *identificación* porque, según la ley federal, la identificación concede a los estudiantes derechos legales que no comparten otros estudiantes; en este caso son de particular relevancia ciertos derechos al respecto de la evaluación, pero también derechos a servicios que

* Los términos *incidencia* y *prevalencia* se confunden a menudo en el lenguaje popular. La incidencia es el número de nuevos casos de una condición que ocurren en la población durante un periodo específico. La prevalencia es el número de casos presentes en un determinado momento. Por ende, la tasa de incidencia puede ser mucho mayor que la tasa de prevalencia si las condiciones son efímeras. La prevalencia de resfriados entre mis estudiantes durante la primera semana de octubre del año pasado fue muy baja. La tasa de incidencia en el curso del semestre fue bastante alta. Las tasas de identificación mostradas aquí son tasas de prevalencia: la proporción de estudiantes del grupo de edad estipulado que presenta una determinada condición de discapacidad en una fecha dada.

suelen ser muy polémicos debido a su costo. En contraste, mucha gente en el campo de la educación especial sostiene que las incongruencias en la *clasificación* no son importantes, entre otras cosas, porque muchos estudiantes padecen más de una discapacidad y la elección de la discapacidad principal puede ser arbitraria. Su principal argumento es que los servicios otorgados a un estudiante deben basarse en los impedimentos funcionales para el aprendizaje de cada individuo, no en la clasificación amplia en que lo ubica su discapacidad. Dos estudiantes que según la clasificación presentan diferentes discapacidades primarias pueden necesitar los mismos servicios, mientras que otros dos con la misma clasificación tal vez requieran servicios diferentes.

No estoy en desacuerdo con este argumento cuando se aplica a la enseñanza y a otros servicios educativos, pero es poco realista cuando se aplica a la evaluación. Tal vez lo ideal sea individualizar por completo la evaluación sin considerar la clasificación, pero muchas veces no es posible. Las categorías de discapacidad son importantes para los propósitos de la evaluación y las incongruencias en la clasificación –a menudo disparatadas– suponen un grave impedimento a la mejora de nuestros métodos para examinar a estudiantes con discapacidades.

Las leyes federales (en particular IDEA y la NCLB, pero también otras más) imponen numerosos requisitos para evaluar a los estudiantes con discapacidades.[3] IDEA exige que los estados establezcan metas de desempeño para los alumnos con discapacidades que sean tan consistentes como sea factible con las metas para los otros estudiantes. También deben incluir a los estudiantes con discapacidades en los programas de evaluación, a nivel del estado y del distrito, que se emplean para evaluar a otros estudiantes, "con las adecuaciones apropiadas donde sea necesario". Los estados también deben poner en práctica "evaluaciones alternativas" para el número relativamente pequeño de estudiantes que no pueden participar en las

evaluaciones educativas generales debido a discapacidades graves.[4] Muchas de las decisiones acerca de la educación y la evaluación de cada niño deben ser tomadas por equipos de un programa individualizado de educación (PIE) en el que participan los padres, los educadores y otros profesionales pertinentes así como, cuando es apropiado, el estudiante. Por lo general, los estados proporcionan a estos equipos directrices para el uso de adecuaciones a la evaluación; en algunos casos prohíben explícitamente algunas de ellas, pero las decisiones son tomadas por el equipo del PIE.

La ley NCLB acepta explícitamente el marco proporcionado por IDEA y se basa en él. Requiere que 95 por ciento de los estudiantes con discapacidades sean examinados, que su desempeño se reporte por separado cuando sean lo bastante numerosos para permitir resultados suficientemente confiables, y que se responsabilice a las escuelas de su progreso. También exige que se hagan "adaptaciones y adecuaciones razonables para los estudiantes con discapacidades".[5] Como veremos, las regulaciones para poner en práctica las disposiciones de la mencionada ley en la evaluación de los estudiantes con discapacidades son draconianas.

La clave de uno de los problemas más difíciles que surgen de la evaluación de los estudiantes con discapacidades es la necesidad de "adecuaciones apropiadas" y de "adaptaciones y adecuaciones razonables", un lenguaje que se repite en las regulaciones de muchos estados. Esos términos se emplean de manera poco consistente, pero seguiré la convención más común y utilizaré "adecuación" para referirme a los cambios en la evaluación que no incluyen alteraciones directas del contenido examinado y que no pretenden cambiar lo que mide la prueba. Estos pueden incluir modificaciones en la presentación de la prueba (como proporcionar una versión en braille a un estudiante ciego), en el contexto o en otros aspectos de la aplicación (permitir que un estudiante presente la prueba en un salón sin otros estudiantes o con pausas

más frecuentes), o en el modo permitido de respuesta (por ejemplo, permitir que un estudiante con un impedimento ortopédico dicte sus respuestas en lugar de escribirlas).

El requisito de usar "adecuaciones apropiadas" parece bastante inocuo; después de todo, difícilmente se querrían adecuaciones inapropiadas. Pero ¿qué hace que los cambios en la evaluación sean "apropiados" y "razonables"? Esta resulta ser una pregunta de extraordinaria dificultad, una de las más desconcertantes en el campo actual de la medición. Para abordarla es necesario empezar con el propósito de las adecuaciones.

Aunque las adecuaciones constituyen una violación deliberada de la estandarización, comparten la importante meta de mejorar la validez de las conclusiones basadas en los resultados obtenidos en las pruebas. Por lo regular, estandarizamos las evaluaciones para eliminar las fuentes engañosas de variaciones en los resultados. Por ejemplo, si yo permitiera que mis estudiantes consulten sus notas cuando presentan un examen y usted no, las calificaciones de nuestros alumnos no serían comparables. Las notas de mis discípulos se verían aumentadas por la mayor indulgencia de mis reglas, independientemente de su nivel real de logro. Por lo tanto, en la mayoría de los casos una comparación basada en una prueba estandarizada proporciona una base más sólida para la comparación que una que no se estandarizó.

No obstante, los resultados que obtienen los estudiantes con ciertas discapacidades pueden ser engañosamente bajos en una prueba aplicada de manera estandarizada. El ejemplo más claro es el de los alumnos con discapacidades visuales. Si a una alumna no le resulta fácil leer un texto, su resultado en una prueba presentada en una forma impresa estándar obviamente será menor de lo que merece su competencia. Las adecuaciones tienen el propósito de compensar impedimentos como este para nivelar el campo de juego –hacer que el resultado obtenido por un estudiante con una

discapacidad sea comparable al mismo resultado obtenido por otro alumno evaluado en condiciones estándar. En el caso de la estudiante con una grave discapacidad visual, eso requeriría la presentación de la prueba en alguna forma distinta al impreso estándar (con letras grandes, en braille o de manera oral).

La metáfora de la adecuación que sugerí al panel del Consejo Nacional de Investigación al que me referí antes fue la de una lente correctiva. Digamos que usted quiere evaluar la competencia de un estudiante en álgebra, por lo que aplica una prueba estandarizada de esa materia. Piense en la prueba como una regla vertical. Los estudiantes con mayor competencia deberían obtener resultados más altos que los colocaran a mayor altura sobre la regla. La prueba proporcionará una estimación de dónde debería ubicarse el marcador de la altura de cada estudiante. Esta estimación resultará confusa debido al error de medición, aunque para la mayoría de los estudiantes, la confusión se extenderá por igual en ambas direcciones, y si los examina en varias ocasiones y saca un promedio, se ajustará gradualmente en la estimación correcta. Pero ¿qué sucede con un estudiante que tiene una discapacidad visual y sólo puede leer los materiales de la prueba con mucha lentitud y con mayor presión? La estimación que usted haga de su competencia no sólo será confusa, también presentará un sesgo que la hará descender más de lo que merece su verdadera competencia. Si lo examinara de manera repetida, podría reducir la confusión pero enfocaría la atención en el resultado equivocado.

Las adecuaciones ideales funcionarían como una lente correctiva, compensando los impedimentos para el desempeño relacionados con la discapacidad y elevando su estimación de la competencia del estudiante al nivel correcto. Esto haría que los resultados obtenidos por estudiantes con discapacidades tuvieran un *significado comparable* a los resultados obtenidos por otros alumnos. Digamos que dos estudiantes, uno con una grave discapacidad

visual que no recibió adecuación y otro sin discapacidad, obtienen una puntuación de 30 (de un máximo de 36) en la parte de matemáticas de la prueba de admisión a la universidad ACT. Sin más información, el funcionario de la oficina de admisiones que recibe esas dos puntuaciones podría inferir que los dos alumnos mostraron una maestría comparable del contenido de matemáticas y de las habilidades medidas por el examen. Esto sería erróneo; sin adecuaciones, el significado de esas dos puntuaciones de 30 no es comparable. En condiciones ideales lo sería si se hicieran las adecuaciones.

Por lo tanto, el propósito de las adecuaciones no es ayudar a los alumnos a obtener mejores puntuaciones sino *ayudarlos a obtener resultados que sean tan buenos como merece su verdadera competencia*, no mayores. En otras palabras, su propósito no es aumentar los resultados sino mejorar la validez. Sin embargo, esta distinción no siempre se reconoce. Hace algunos años, la directora de educación especial de un estado me dijo que las adecuaciones deberían ofrecerse a todos los estudiantes, no sólo a los que presentaban discapacidades. Con el temor (razonable) de conocer la respuesta, le pregunté por qué, a lo que respondió "Les iría mejor". Esto es un poco como decir que a los niños que no quieren ir un día a la escuela se les debería dar la oportunidad de meter el termómetro en agua caliente antes de pasárselo a sus padres.

El mayor problema en la evaluación de los estudiantes con discapacidades puede ser que muchas veces no sabemos qué adecuaciones compensarán el sesgo ocasionado por la discapacidad sin proporcionarles una ventaja injusta. Peor todavía, existen casos en que ni siquiera es factible diseñar adecuaciones que sean del todo convenientes. Utilizaré dos casos como ejemplo. Para las discapacidades visuales, que son muy raras, tengo una idea bastante buena, aunque no perfecta, de las adecuaciones que deberíamos proporcionar. Las discapacidades para el aprendizaje, que son mucho más comunes, plantean dificultades mucho mayores.

Hace varios años tuve una alumna que padecía un grave problema visual, acromatopsia congénita. En la gente con visión normal los conos tienen la función de ver en condiciones de luz brillante mientras que los bastones se encargan de la visión en condiciones con poca luz. La gente con acromatopsia carece de visión normal de los conos por lo que depende de los bastones en todas las condiciones de iluminación. Las consecuencias son una pobre agudeza visual, la incapacidad para adaptarse a la luz brillante (los bastones se saturan con niveles de iluminación relativamente bajos) y diversos grados de ceguera al color. La condición de mi alumna era lo bastante grave para identificarla como legalmente ciega, aunque su falta de visión no era absoluta.

Gracias a sus sugerencias pude proporcionarle algunas adecuaciones que la ayudaron a funcionar bien en la clase. Disminuí la iluminación, le reservé un asiento que la colocaba de espalda a la ventana y le proporcionaba la mejor visión posible de la pantalla sobre la que proyectaba las transparencias, también le daba materiales impresos con letras mucho más grandes. Trataba de no escribir o dibujar en el pizarrón porque ella no podía leer lo que escribía, pero cuando tenía que hacerlo, un asistente de enseñanza copiaba lo que yo escribía y se lo entregaba. Con esas adecuaciones pudo seguir muy bien mis clases.

Sin embargo, mis exámenes, que eran estandarizados, representaban un problema para ella. Yo aplicaba las pruebas en forma impresa y los alumnos tecleaban sus respuestas directamente en las computadoras en un laboratorio con iluminación brillante. A ella no le resultaba sencillo leer las letras estándar, sobre todo en condiciones de luz brillante, y los caracteres en la pantalla de la computadora eran demasiado pequeños para que pudiera leerlos. En esas condiciones estándar su desempeño en el examen habría sido engañosamente bajo.

En respuesta, le aplicaba los exámenes en el salón contiguo con una iluminación muy reducida. Le proporcioné una computadora con una pantalla más grande y un programa que le permitía cambiar el tamaño de los caracteres. Las preguntas del examen se cargaban en la computadora de modo que pudiera leerlas usando el tamaño de letra que fuera mejor para ella. Todo lo anterior violaba la estandarización, pero en este caso, el efecto fue sin lugar a dudas que su resultado (que al final fue muy alto) representara de manera más exacta su verdadero dominio del contenido del curso.

Sin embargo, el final feliz de este relato no debe dejarnos optimistas con respecto a la situación general. A diferencia de lo que sucede en la mayor parte de los casos, varias razones nos permitieron hacer adecuaciones eficaces para esta estudiante. Nos ayudó el nivel de recursos que pudimos utilizar, el cual puede ser igualado por pocas escuelas. No obstante, otras razones de nuestro éxito nada tienen que ver con los recursos y son más importantes para la perspectiva general. La discapacidad de la estudiante tenía cuatro características que facilitaron el diseño de las adecuaciones apropiadas.

Primero: la discapacidad de la alumna (tanto el hecho de que tenía una discapacidad como la *discapacidad específica que padecía*) era inequívoca. Segundo: aunque su discapacidad era poco común, era bien conocida y sus consecuencias para el desempeño en una prueba aplicada en condiciones estandarizadas parecían claras. Sus síntomas se ajustaban como anillo al dedo a la descripción estándar del síndrome y la clasificación de su discapacidad por sí sola era suficiente para indicar varias adecuaciones apropiadas, como usar caracteres grandes y disminuir la iluminación. Esto no es distinto de lo que ocurre a menudo en la medicina: uno va al médico por los síntomas y el diagnóstico suele ser la clave para un tratamiento exitoso.

Tercero: –y no puede exagerarse la importancia de este punto– el impedimento que enfrentaba esta estudiante, su falta de agudeza visual, *no se relacionaba con el contenido ni con las habilidades que la*

prueba estaba diseñada para medir. En la fea jerga del oficio, las barreras que enfrentaba esta chica para tener un buen desempeño en el examen eran "irrelevantes para el constructo". Por lo tanto, los efectos de la discapacidad sobre su desempeño en la prueba estandarizada constituían un claro sesgo: si se le aplicaba el examen en la forma estándar su resultado implicaría un nivel de dominio inferior al que en realidad había alcanzado. Si pudiésemos encontrar una adecuación que no hiciera otra cosa sino compensar este impedimento, la validez aumentaría.

Y esto apunta al cuarto y último factor que operó a nuestro favor: parecía razonablemente clara la manera de diseñar las adecuaciones prácticas para aliviar el sesgo sin sesgar sus resultados en la otra dirección. Por ejemplo, no había razón para esperar que disminuir la iluminación o aumentar el tamaño de los caracteres le hubieran conferido una ventaja injusta.

Pero incluso en el caso de las discapacidades visuales, con las que es mucho más sencillo lidiar con adecuaciones que en el caso de otras discapacidades, el resultado podría no ser tan bueno como con esta alumna porque la compensación del impedimento podría ser insuficiente o excesiva. Hace algunos años, un estudiante ciego indicó a los investigadores del Servicio de Evaluación Educativa (*Education Testing Service*-ETS) que en el caso de ciertos tipos de exámenes, el sólo hecho de presentar el texto en braille no nivelaba del todo el campo de juego. No es inusual que los estudiantes deban regresar al texto de un reactivo, sobre todo si es largo y complejo, para extraer información específica. Por ejemplo, una tarea compleja de matemáticas podría requerir que se regresara de manera repetida al texto para buscar datos numéricos o para extraer información de una gráfica. Según el reclamo de este alumno, esto se lleva más tiempo para los estudiantes que leen braille porque ellos no pueden dar una ojeada rápida y tienen que revisar con mayor lentitud todo el reactivo o buena parte de él

para relocalizar la información. El estudiante sostenía que esto los coloca en desventaja, sobre todo si no reciben tiempo adicional.[6]

Yo no leo en braille y no puedo dar fe de la pertinencia de la preocupación de este alumno, pero el problema de las ojeadas rápidas afecta a muchos buenos estudiantes con discapacidades o con un dominio limitado del inglés y es un buen ejemplo del tipo de sutilezas que conlleva el diseño de adecuaciones eficaces. Aunque puede no ser obvio, quienes leen el inglés con fluidez por lo regular no lo hacen letra por letra sino que suelen percibir de un vistazo partes de palabras o palabras enteras. (De hecho, para leer el inglés con fluidez no es práctico hacerlo letra por letra porque la pronunciación incluso de secuencias de letras varía de una palabra a otra; por ejemplo, compare las últimas cuatro letras de *through* con e*nough*.) Una indicación de ello es que en lugar de estancarse al deletrear incorrectamente las palabras, como sucedería si leyeran letra por letra, quienes leen con soltura las entienden fácilmente al sustituirlas por las letras correctas con rapidez y a veces sin tener conciencia de ello. Por ejemplo, por lo menos desde 2003 ha circulado en internet el texto de la figura 12.1. A pesar de la mezcolanza

Figura 12.1. Texto con letras mezcladas en cada palabra

De aeucdro a una ivecstnigóain en la Uivsreniadd de Cmdgbraie, no ipmotra en qué odern etésn las ltears de una plaraba, la úicna csoa ipomrttnae es que la pemirra y la útmila lreta se euncetnren en la psoóicin crrtocea. Las dámes ltares pdeuen ser un cmpolteo dseoredn y aún así uestd prodá lelears sin nniúgn pmorlbea. Etso se dbee a que el crebero haumno no lee cdaa ltrea de fmroa aaislda, snio la parabla cmoo un todo.

Figura 12.2. Texto invertido pero corregido

De acuerdo a una investigación en la Universidad de Cambridge, no importa en qué orden estén las letras de una palabra, la única cosa importante es que la primera y la última letra se encuentren en la posición correcta. Las demás letras pueden ser un completo desorden y aún así usted podrá leerlas sin ningún problema. Esto se debe a que el cerebro humano no lee cada letra de forma aislada, sino la palabra como un todo.

de letras, la lectura de este pasaje resultará trivialmente fácil para la mayoría de los angloparlantes nativos. Existe mucho desacuerdo respecto a si, como afirma el pasaje, pueden hacerlo gracias a la primera y la última letra de las palabras mal deletreadas. Pero aun así queda claro que los lectores nativos pueden percibir palabras enteras y compensan con mucha rapidez los errores del texto.[7]

Ahora, para hacer una simulación artificial de lo que enfrentan quienes no leen con soltura, vea la figura 12.2, en la cual se corrigieron todos los errores del texto pero éste se invirtió horizontalmente. Si usted es como yo, le resultará muy difícil leer este texto correcto pero invertido y sólo podrá hacerlo con mucha lentitud. Quienes hayan aprendido un idioma que utiliza un alfabeto distinto al latino que se emplea en el inglés tal vez han experimentado una especie de *déjà vu* cuando trataron de leerlo: la lectura es similar a la tarea de leer una ortografía desconocida. Cuando las *letras* se invierten su lectura resulta moderadamente difícil, pero ninguna de las *palabras* invertidas puede reconocerse de un vistazo, por lo que nos vemos obligados a leer el texto letra por letra para luego reunir las letras en palabras. Esto es parecido al proceso usado por muchos lectores inexpertos y por algunos niños mayores con dificultades de lectura. También puede ser similar al proceso de búsqueda descrito al ETS por el estudiante que leía en braille.

Una solución evidente sería otorgar tiempo adicional a los estudiantes que tuvieran que leer en braille, de hecho esa es la adecuación más común que se ofrece en los sistemas que aplican exámenes con límites de tiempo. Pero ¿cuánto tiempo más debería concederse? Si se ofrece demasiado tiempo adicional se corre el riesgo de sobrecompensar (crear una ventaja injusta) en lugar de sólo nivelar el campo de juego. En un caso reciente, un estudiante que tenía que presentar el examen estatal de abogacía, cuya aplicación se lleva por lo regular 18 horas a lo largo de tres días, solicitó y obtuvo casi cinco veces esa cantidad de tiempo

extendida a lo largo de casi dos semanas. ¿Es cinco veces la cantidad normal suficiente para compensar los efectos de la discapacidad sin conferir una ventaja injusta? ¿Diez veces?

Un estudio realizado hace dos décadas sobre los resultados obtenidos por estudiantes con discapacidades en las pruebas SAT y GRE demostró que no se trata de una preocupación abstracta:

> A excepción de los resultados de estudiantes con problemas auditivos, las puntuaciones obtenidas en la prueba SAT de… [aplicaciones con adecuaciones] tienen una fuerte tendencia a predecir en exceso el desempeño en la universidad de los estudiantes con discapacidades. [Es decir, sus resultados en la prueba son superiores a lo que su desempeño posterior en la universidad consideraría certero]. Este efecto es mayor para los estudiantes con problemas de aprendizaje que reciben puntuaciones relativamente altas… Una explicación posible... es… la política de conceder tiempo ilimitado a las personas que presentan la prueba en condiciones especiales… Existe alguna indicación de que la ganancia ocurre para los estudiantes cuya discapacidad hace necesario el tiempo adicional… pero también hay una indicación de que estudiantes más capaces están recibiendo cantidades mayores de tiempo.[8]

A pesar de las dificultades para obtener las adecuaciones correctas para los estudiantes con discapacidades visuales, este tipo de impedimento es uno en que resulta más sencillo hacer adecuaciones. Suele ser mucho más difícil elegir las adecuaciones que generarán una estimación razonablemente no sesgada del nivel de competencia. En primer lugar, a menudo no quedan claras las discapacidades específicas que presentan los estudiantes, como lo demuestran los caóticos patrones en las clasificaciones de un estado a otro y entre distritos y escuelas de muchos estados. E incluso cuando la discapacidad es clara, tal vez no lo sean los impedimentos que ocasiona.

Sin embargo, la dificultad fundamental es que, en algunos casos, los impedimentos ocasionados por la discapacidad son directamente relevantes para el conocimiento y las habilidades que la prueba pretende medir. En esos casos, compensar esas barreras con adecuaciones puede crear una ventaja injusta y un sesgo potencial de los resultados en la otra dirección. No se trata sólo de un problema técnico difícil; en algunos casos puede no ser del todo factible crear las adecuaciones razonables exigidas por la ley. En ocasiones puede suceder lo mismo al examinar a estudiantes cuyo dominio del inglés es limitado. Esta es la razón principal del pesimismo con que inicié este capítulo.

El problema de las discapacidades relevantes para el constructo tal vez sea más claro en el caso de dificultades específicas para el aprendizaje, las cuales constituyen la mayor categoría de las discapacidades y dan cuenta de cerca de la mitad de todos los estudiantes que son atendidos bajo la ley IDEA. La discapacidad de aprendizaje más común, la dislexia, interfiere con la capacidad lectora de los estudiantes. "Leer" puede significar varias cosas, pero para este propósito vamos a usar una definición sencilla: la capacidad para descodificar e inferir el significado de textos impresos tal como suelen presentarse (por ejemplo, en libros y periódicos). La dislexia interfiere con el proceso implicado en la descodificación, como la diferenciación de fonemas, pero no en los procesos cognoscitivos de orden superior que participan en la lectura, como sacar inferencias del texto.

Ahora bien, para hacer al ejemplo tan extremo como sea posible, considere las pruebas de lectura, que son exigidas por la ley NCLB y que en cualquier forma ya desde antes se aplicaban de manera generalizada. Estas pruebas plantean el irritante problema de no permitirnos separar los impedimentos causados por la discapacidad del constructo que tratamos de medir. Sin las adecuaciones, mi alumna con acromatopsia habría tenido un pobre

desempeño en mi examen por su incapacidad para leer las letras pequeñas, pero esa discapacidad era del todo irrelevante para los constructos que yo quería medir con mi instrumento. La situación cambia cuando un estudiante disléxico presenta una prueba de lectura. Su dislexia obstaculiza su capacidad de leer bien la prueba, pero eso es justamente lo que pretendemos medir con la prueba. Es posible que haya aspectos de la tarea de lectura que el estudiante desempeña mejor de lo que indica su resultado (por ejemplo, la debilidad de su descodificación puede ocultar una sólida capacidad para sacar inferencias de pasajes del texto), pero aun así su competencia general en la lectura es pobre porque no puede descodificar bien, y no hay una forma obvia de usar apropiadamente las adecuaciones. Podríamos leerle la prueba o presentársela en una grabación, lo que nos permitiría sortear la dislexia, pero esto cambiaría en lo esencial lo que mide la prueba, la cual ya no mediría "la capacidad para descodificar e inferir el significado del texto impreso tal como suele presentarse". Mediría otra cosa, tal vez "la capacidad para entender y sacar inferencias del lenguaje oral". Si usted fuera un empleador que busca quien ocupe un puesto que requiere de mucha lectura, ¿qué resultado consideraría una base más válida para evaluar a este estudiante: el resultado obtenido en condiciones estandarizadas, donde el estudiante tuvo que leer el texto, o el resultado obtenido con esta adecuación, que sólo requería que el estudiante comprendiera el lenguaje oral?

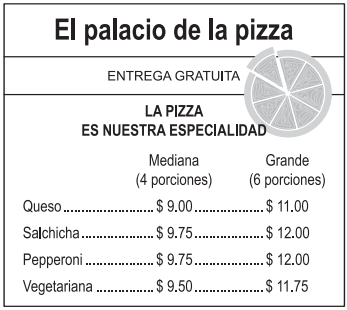
Usted podría plantear que este ejemplo es demasiado extremo, pero de hecho los educadores y los políticos discuten ahora sobre cuál es la mejor manera de aplicar las pruebas de lectura a los estudiantes disléxicos y, en cualquier caso, el problema no se limita a las pruebas de lectura. El ejemplo usado por el panel del Consejo Nacional de Investigación que se mencionó antes para ilustrar este problema era el de las pruebas de matemáticas. En la actualidad, muchas pruebas de matemáticas se esfuerzan por incluir problemas

realistas del tipo que los estudiantes pueden encontrar fuera de la escuela. Muchas de esas pruebas conllevan una buena cantidad de lectura; algunas requieren que los estudiantes escriban las explicaciones de sus respuestas. En la figura 12.3 se muestra un ejemplo de dichos reactivos, tomado de una prueba de matemáticas de secundaria.

Es claro que entre más lectura y escritura requiera una prueba de matemáticas, más probable es que los resultados de los estu-

■ **Figura 12.3.** Un reactivo de matemáticas de octavo grado que requiere considerable lectura y escritura. Tomado de la evaluación MCAS de Massachusetts aplicada en 2000

38. El comité de planeación de la Secundaria Lane está organizando una fiesta de pizza para sus 127 estudiantes de octavo grado. Obtuvieron este menú de El palacio de la pizza.



El comité de planeación aplicó una encuesta a una muestra aleatoria de 26 estudiantes de octavo grado en que les preguntó, "¿Qué tipo de pizza quieres?". Esto es lo que encontró.

Tipo favorito de pizza				
Tipo de pizza	Queso	Salchicha	Pepperoni	Vegetariana
Número de estudiantes	7	3	9	7

El comité tiene un presupuesto de $300 para la pizza. ¿Qué tipos y tamaños de pizza podría ordenar para que cada uno de sus 127 estudiantes pueda tener su tipo favorito de pizza?

a. Explique cómo utilizó los resultados de la encuesta para decidir qué pizzas ordenar.

b. Muestre o describa los cálculos que se necesitan para estar seguros de que habrá suficiente pizza para los 127 estudiantes.

c. Muestre o describa los cálculos necesarios para estar seguros de que el costo de las pizzas será igual o menor a $300.

No es necesario que encuentre la forma más barata de comprar suficiente pizza. Sólo tiene que asegurarse de que el costo total es igual o menor a $300.

diantes disléxicos se vean afectados por sus discapacidades. Pero ¿este efecto adverso es un sesgo que debería ser compensado con una adecuación o es de hecho un indicador realista de baja competencia? Según el panel del Consejo Nacional de Investigación, la respuesta es que eso depende de la *inferencia específica* que se sustente en los resultados, es decir, depende de a qué quiere referirse con "competencia en matemáticas". Por un lado, si estuviera usando los resultados para estimar habilidades como la facilidad para el cálculo, es claro que el desempeño de los estudiantes disléxicos en las pruebas de matemáticas que requieren mucha lectura tendría un sesgo hacia abajo porque sus dificultades para la lectura les harían difícil demostrar sus habilidades para el cálculo. Por otro lado, si estuviera usando los resultados para estimar qué tan bien pueden aplicar los estudiantes esas habilidades a problemas reales del tipo que pueden encontrar más tarde, incluyendo problemas de matemáticas insertados en texto, el sesgo de los resultados sería menor. Si se diseñara una adecuación que compensara por completo las dificultades para la lectura de esos estudiantes, estos obtendrían resultados engañosamente altos porque se habría sobreestimado su capacidad para aplicar las habilidades matemáticas a problemas del mundo real presentados en texto. Los participantes en el panel del Consejo Nacional de Investigación vieron esto como otro ejemplo de los compromisos involucrados en la evaluación: consideraron que hay buenas razones para incluir esas tareas realistas en las pruebas de matemáticas, pero reconocieron que el costo sería que resultarían difíciles para los estudiantes con discapacidades.

Este problema era el meollo del caso relacionado con las discapacidades sobre el que decidió la Suprema Corte en 1979. A una mujer se le negó la admisión a un programa de entrenamiento en enfermería debido a una severa discapacidad auditiva. La razón ofrecida para rechazarla fue que era incapaz de entender el habla sin leer los labios y que eso hacía imposible que tuviera un

desempeño adecuado en el programa de entrenamiento o que brindara una atención segura a los pacientes. La mujer entabló una demanda bajo la Sección 504 de la Ley de Rehabilitación de 1973, otra ley federal que protege los derechos de los discapacitados y que también repercute en la educación pública desde el nivel básico al medio superior. La Suprema Corte falló a favor de la escuela con el argumento de que la discapacidad de la demandante minaba su cualificación para ejercer como enfermera. Si a pesar de su discapacidad se le hubiera considerado "por lo demás cualificada" (el lenguaje legal) habría ganado la demanda.[9] En la jerga de la evaluación, si las consecuencias de su discapacidad hubieran sido irrelevantes para el constructo (su funcionamiento como enfermera), habría tenido un mejor caso, pero de hecho eran directamente relevantes para él.

Los datos de prevalencia anotados en la tabla 12.1 dejan claro que son muchos los casos en que las adecuaciones apropiadas son ambiguas y que los casos bien definidos son una extraña excepción. Por ejemplo, en el año escolar 1999-2000 los estudiantes con discapacidades visuales constituían menos de la mitad del 1 por ciento de los estudiantes con discapacidades y 5 centésimas del 1 por ciento (cinco estudiantes por cada 10 000) de la población en edad escolar. Los estudiantes clasificados con problemas ortopédicos (otro grupo para el que las adecuaciones son razonablemente claras) ascendían a 14 por cada 10 000 estudiantes. Los números grandes se encuentran en los grupos que plantean problemas mucho más difíciles, en particular, los estudiantes con dificultades de aprendizaje.

Imagine entonces que es miembro de un equipo de un programa individualizado de educación responsable de elegir las adecuaciones para la evaluación apropiada de un estudiante con discapacidad. Supongamos que tiene un alumno con una discapacidad que resulta problemática para la evaluación, como un

problema de aprendizaje. ¿Dónde encontraría orientación basada en evidencia empírica sólida para determinar la manera de examinar a este niño?

Un primer paso lógico sería recurrir a las directrices de su estado para evaluar a los estudiantes. No sé cómo suelen hacer esto los equipos de los programas individualizados de educación, aunque los datos de los educadores relacionados con la asignación de adecuaciones sugieren que muchos de ellos no prestan mucha atención a las directrices. De hecho, la ley actual ofrece un incentivo para ignorar cualquier restricción implicada por las directrices estatales porque el hecho de no proporcionar adecuaciones suficientes para mejorar los resultados puede tener costos considerables (como no lograr el progreso anual adecuado según la ley NCLB) mientras que los riesgos por proporcionar un exceso de adecuaciones son menores. No obstante, suponga que usted y sus compañeros de equipo revisan cuidadosamente las directrices.

En muchos casos no encontrará su respuesta. Aunque algunos estados han producido directrices informativas y escritas con cuidado, estas son insuficientes para identificar las adecuaciones apropiadas para muchos estudiantes. En primer lugar, dichas directrices cambian de manera sistemática de un estado a otro, e incluso existen casos en que un estado permite en forma explícita una adecuación que otro prohíbe expresamente.[10] Esas directrices conflictivas no pueden ser todas correctas, de modo que incluso si sigue las directrices de un determinado estado, podría suceder que se encuentra en un estado cuyo consejo es erróneo. Además las directrices estatales suelen ser demasiado generales para resolver su problema. Por ejemplo, muchos advierten a los equipos de los programas individualizados de educación que no deben usar adecuaciones que cambien el significado del resultado obtenido en la prueba o que minen la validez, pero por lo general no explican lo que esto significa en la práctica –por ejemplo, no indican qué

adecuaciones específicas cumplirían esta meta en el caso de estudiantes con discapacidades específicas.

Usted podría volverse ambicioso y decidir que quiere explorar por sí mismo la investigación relevante. Como alguien que ha llevado a cabo parte de ese trabajo, puedo asegurarle que se sentiría desilusionado. Los estudios no son muchos, algunos no son muy buenos y dejan sin responder la mayoría de las preguntas importantes. De hecho, el débil estado de las publicaciones científicas es uno de los motivos por los que las directrices estatales no son más útiles.

Existen varias razones por las cuales la investigación deja tanto que desear. Una es el simple hecho de que hasta hace muy poco tiempo no eran muchas las personas que trabajaban en este campo. Otra es que puede ser políticamente difícil realizar ciertos tipos de trabajo de alta calidad en esta área. Por ejemplo, a finales de la década de los noventa obtuve dos veces, la autorización del superintendente de un estado para realizar verdaderas pruebas experimentales de los efectos de las adecuaciones. Para este propósito, los experimentos verdaderos, con sujetos asignados al azar a los tratamientos, son la regla de oro pero son muy pocos los que se han realizado para evaluar las adecuaciones. Ambos estudios fueron desautorizados por directores de nivel medio del departamento de educación del estado mucho después de que el trabajo había empezado. La segunda vez confronté a uno de los directores e insistí en recibir una explicación. El estudio, según su explicación, implicaba un riesgo político demasiado grande.

La investigación en este campo enfrenta tres problemas adicionales que son más importantes. Uno es la caótica clasificación de las discapacidades de los estudiantes. Cualesquiera que sean los argumentos contra el uso de las clasificaciones al elegir servicios para los estudiantes, es indiscutible que son importantes para la evaluación. Sería absurdo, por ejemplo, ofrecer a un estudiante con limitaciones ortopédicas las adecuaciones que yo le brindé a

mi alumna con discapacidad visual. Como se advirtió en el estudio del Consejo Nacional de Investigación que mencioné antes:

> Para diseñar una adecuación que incremente la validez… de los resultados de los estudiantes con discapacidad, primero se deben identificar la naturaleza y la gravedad de las distorsiones que compensará la adecuación. Esas distorsiones dependen de la discapacidad… Como las clasificaciones de la discapacidad nos dicen quién puede tener características funcionales subyacentes vinculadas con distorsiones potenciales de los resultados, las ambigüedades o incongruencias en la clasificación de estudiantes con discapacidad tienen implicaciones graves para las evaluaciones… Si la clasificación de una discapacidad es incorrecta o imprecisa, será difícil determinar si las adecuaciones elegidas son válidas.[11]

Digamos que quiere determinar la eficacia de un nuevo medicamento para la tuberculosis resistente a múltiples medicamentos. Obviamente tendría que empezar con personas que padecieran esta forma de tuberculosis y debería examinar los efectos del medicamento en los síntomas causados por la enfermedad. De igual modo, si queremos examinar la eficacia de una presentación con caracteres grandes para alumnos con discapacidad visual, tenemos que empezar con estudiantes que las presenten. Dado que los criterios para el diagnóstico de la mayor parte de las discapacidades son muy poco claros y que las clasificaciones resultantes son demasiado incongruentes, es difícil obtener muestras razonables y generalizar a los estudiantes con discapacidades similares en la población general.

Otra dificultad es que rara vez disponemos de un *criterio*, es decir, una medida en la que podamos confiar lo suficiente para usarla como estándar para evaluar los efectos de varias adecuaciones. Por ejemplo, digamos que queremos averiguar qué tan bien compensan

dos cantidades de tiempo adicional (el tiempo normal más la mitad y el doble del tiempo normal) el sesgo producido por una discapacidad específica. De modo que aplicamos la prueba de tres maneras a tres grupos de estudiantes similares con esta discapacidad: a un grupo se le da la cantidad estándar de tiempo, otro recibe una y media veces el tiempo normal y a un tercer grupo se le otorga el doble del tiempo. Encontramos luego que el primer grupo recibió los resultados más bajos y el tercero los más altos. ¿Cuál de los tres conjuntos de resultados es el más exacto? ¿Con qué debería compararlos para averiguarlo? En la mayor parte de las evaluaciones aplicadas desde la educación básica a la media superior carecemos de un estándar digno de confianza para hacer la comparación.*

Por último, está el problema particularmente irritante de las discapacidades que crean impedimentos relacionados con el constructo medido por la prueba, por ejemplo, como cuando necesitamos juzgar el logro de los estudiantes con dislexia. Este es un problema tanto lógico como técnico y en esos casos enfrentamos una ambigüedad fundamental en nuestra interpretación de los resultados. Dado este problema lógico, parece poco probable que la investigación señale una manera de obtener resultados de esos estudiantes que sean del todo comparables con los de otros alumnos. Una meta más realista, aunque menos satisfactoria, sería mejorar lo que conocemos acerca de esos estudiantes, incluso si la información recogida no es del todo comparable con la que tenemos de otros. Considere de nuevo el problema de evaluar en matemáticas a estudiantes con dislexia cuando la evaluación incluye intencionalmente habilidades matemáticas en contextos "del

* Una excepción son las pruebas de admisión a la universidad, diseñadas para predecir el desempeño en la educación superior, por lo que podemos usar como criterio este desempeño (por ejemplo, las calificaciones obtenidas en el primer año). Esto se hizo en el estudio del Servicio de Evaluación Educativa (ETS) mencionado unas páginas antes.

mundo real" que implican leer y escribir. Sería factible desarrollar una evaluación y un sistema de adecuaciones que permitieran sacar conclusiones como "este estudiante tiene sólidas habilidades de cálculo pero no puede aplicarlas a problemas que incluyan texto". Esta sería una inferencia muy útil, pero no nos permitiría decir "un resultado de 143 de este estudiante discapacitado indica el mismo nivel de competencia en matemáticas –tal como la definimos para incluir aplicaciones a problemas escritos– que la misma puntuación obtenida por otros estudiantes".

Otro problema que enfrentamos al evaluar a estudiantes con discapacidades es decidir cómo examinar a los alumnos con desempeño muy bajo. Por supuesto, no todos los estudiantes con discapacidades tienen un mal desempeño pero, en general, estos alumnos están representados de manera desproporcionada en el extremo inferior de la distribución.

¿Cómo debería evaluarse a los estudiantes con muy bajo desempeño, discapacitados o no? Si lo único que interesa es la medición exacta –y ese es un "si" muy grande porque la evaluación tiene muchas metas–, la respuesta es clara. La prueba hipotética de vocabulario descrita en el capítulo 2 ofrece un primer indicio. Una prueba demasiado difícil proporciona muy poca información sobre el nivel de desempeño de un examinado. De enfrentar una prueba de vocabulario que incluyera palabras como *silícula*, la mayoría de nosotros habríamos errado prácticamente en todos los reactivos, lo cual habría producido un "efecto de piso": un montón de resultados apilados cerca de la puntuación más baja posible. Tal resultado nos habría dicho que ninguno de nosotros conoce el significado de *silícula*, pero no nos habría dado información acerca de dónde se ubica cada uno de nosotros en el continuo de las habilidades de vocabulario.

En este caso participan dos principios diferentes. El primero es el simple hecho de que una prueba de logro debe medir el

contenido que los examinados están estudiando en la actualidad o se espera que estudien. Si las clases de matemáticas de un alumno de secundaria con una grave discapacidad cognitiva se enfocaron en los conceptos de número y una aritmética muy básica, sería absurdo aplicarle una prueba en la que deba resolver ecuaciones lineales. En cambio, tenemos que examinar a estos estudiantes de manera diferente a los otros, por ejemplo aplicándoles un examen diseñado para los grados anteriores. Este principio puede parecer obvio, pero es común que se ignore (cada vez con mayor frecuencia, como veremos en un momento).

El segundo, incluso dentro de un dominio de contenido, es que la dificultad de una prueba debe corresponder al nivel de desempeño del estudiante. La razón es la confiabilidad. Si todo lo demás se mantiene constante, hacer que una prueba resulte demasiado difícil o demasiado fácil para una persona incrementa la cantidad de error y disminuye la confiabilidad. En el capítulo 7 expliqué el error estándar de medición, el rango de incertidumbre que rodea al resultado de cualquier prueba. En la práctica, no hay un error estándar único, sino muchos: el margen de error para los estudiantes para los cuales la prueba tiene un grado apropiado de dificultad es menor que para los estudiantes con los resultados más altos y más bajos. Este es un motivo por el que algunas pruebas se aplican por computadora, como es el caso del Examen de Registro de Posgrado (*Graduate Record Examination*-GRE) que presentan quienes quieren ingresar a la mayor parte de las escuelas de posgrado. Este examen es una *prueba adaptativa por computadora* (PAC), en que el desempeño de los estudiantes en los primeros reactivos da lugar a que se les asignen reactivos más sencillos o más difíciles que correspondan mejor con su nivel de desempeño. El resultado es un nivel de confiabilidad más alto porque los alumnos no desperdician tiempo en reactivos que son demasiado fáciles o demasiado difíciles para ellos. Cuando el desequilibrio entre

la competencia de un estudiante y la dificultad de la prueba es grave (como sucede con algunos estudiantes con discapacidades en muchos sistemas actuales de evaluación), disminuye mucho la confiabilidad de los resultados.

Para ser justos con los políticos que recientemente tomaron lo que yo considero malas decisiones acerca de la evaluación de los estudiantes con bajo desempeño (incluyendo algunos con discapacidades), determinar la manera de examinar a esos estudiantes plantea un dilema importante. La medición óptima no es la única preocupación. Cuando se utiliza una prueba para la rendición de cuentas, diseñarla para que sea psicométricamente apropiada para los estudiantes que obtienen bajas puntuaciones puede reducir los incentivos destinados a elevar el logro de esos alumnos. Esta es una preocupación importante toda vez que, por tradición, a muchos estudiantes con discapacidades se les han presentado currículos poco exigentes que implican una grave limitación para sus opciones posteriores de educación y trabajo. La preocupación se agrava por el hecho de que el logro es un continuo. Si uno hace concesiones especiales para estudiantes que están muy por debajo del promedio, ¿qué hay acerca de quienes se desempeñan un poquito mejor? ¿En qué punto deja de ser razonable utilizar un estándar más bajo para los estudiantes? No son preguntas para las que haya una respuesta fácil.

No obstante, las decisiones tomadas bajo la ley NCLB acerca de la evaluación de estudiantes de bajo logro, aunque aún están en evolución y gradualmente se vuelven menos rigurosas, son draconianas. La historia, tal como ha surgido en una serie de regulaciones y reseñas de la reglamentación propuesta (RRP) es barroca y requiere que se distinga entre "evaluaciones alternativas", "estándares alternativos" y "estándares modificados". Como se mencionó antes, la ley IDEA requiere que los estudiantes con discapacidades demasiado graves para presentar la prueba regular, incluso con

adecuaciones, presenten una "evaluación alternativa". La ley no especifica cuántos son esos niños, pero la expectativa general es que son pocos. Se maneja mucho que constituyen alrededor del 2 por ciento de todos los estudiantes, aunque no he podido determinar de dónde salió ese número.

Las primeras reseñas de las reglamentaciones propuestas denominaron "estándares alternativos" a las normas más bajas y permitieron que los distritos y estados los aplicaran sólo a medio punto porcentual de los estudiantes. La RRP afirmó de manera explícita que los "estándares alternativos" son distintos de la "evaluación alternativa" requerida por la IDEA y que algunos alumnos podrían presentar esta última y al mismo tiempo estar sometidos a los estándares del nivel del grado. Originalmente la lógica que limita el uso de los estándares alternativos a la mitad de un punto porcentual también era explícita: los estudiantes con un retardo leve deberían someterse a los estándares del nivel de grado. Sólo los alumnos con discapacidades cognitivas más graves –quienes tienen retardo moderado, severo y profundo– tendrían derecho a los estándares más bajos. La reseña de la reglamentación propuesta los identificó específicamente como los estudiantes cuyo desempeño era inferior a la media en más de tres desviaciones estándar. Cualquier estudiante por arriba del límite del medio punto porcentual evaluado con los estándares alternativos recibiría automáticamente la etiqueta de "no competente" para los propósitos de rendición de cuentas (es decir, para determinar el progreso anual adecuado), lo que en consecuencia pondría al distrito en mayor riesgo de sanciones. La versión final de esas regulaciones, publicadas en diciembre del 2003, incluyó algunos cambios importantes. El límite se elevó a 1 por ciento para permitir variabilidad geográfica en la prevalencia de las discapacidades cognoscitivas. Se eliminó la referencia al retardo mental y a los estudiantes con más de tres desviaciones estándar por debajo de la media, debido en parte a que se

temía que eso diera lugar a una dependencia excesiva de las pruebas de CI para clasificar a los estudiantes. Sin embargo, esto parece haber sido una cuestión de terminología y las regulaciones conservaron el requisito de utilizar estándares regulares para evaluar a los estudiantes con discapacidades cognoscitivas leves.[12]

¿Qué tan draconianos son esos requisitos? Una forma de ponerlos en perspectiva es considerar lo que están estudiando los alumnos con bajos resultados a quienes se tomó en cuenta para cumplir con los estándares regulares. Puede esperarse que los alumnos de octavo grado con discapacidades cognoscitivas leves que están en un programa educativo bien diseñado lean a un nivel de cuarto o quinto grado y que hayan progresado en matemáticas al punto de realizar aritmética simple. Sin embargo, según las previsiones de la ley NCLB, en el lapso de 12 años esos estudiantes deberían alcanzar el estándar competente, que la ley define como un "nivel de logro alto" para el grado que estudia el alumno.[13] No más aritmética simple; ahora estamos hablando de habilidades previas al álgebra y de la aplicación de la aritmética a problemas de complejidad razonable, como el reactivo de matemáticas de octavo grado que se muestra en la figura 12.3.

La información normativa hace aún más clara la severidad de esos requisitos. Digamos que los estados adoptan estándares de competencia de dificultad comparable a los de la Evaluación Nacional del Progreso Educativo, tal como insisten muchos reformadores de la educación. Como mencioné en el capítulo 4, casi una tercera parte de los estudiantes de Japón y Corea no podrían alcanzar el estándar de competencia de la NAEP si se les aplicara dicha evaluación. Por lo tanto, la puesta en práctica de las regulaciones de la ley NCLB implica que al final del periodo de 12 años especificado por dicha ley, los alumnos estadounidenses con retardo leve deberían superar aproximadamente a la tercera parte de los estudiantes de dos de los países con mayores puntuaciones de

todos los que han participado en las evaluaciones internacionales de matemáticas. Este sería el enfoque del optimismo y el trabajo duro (ejemplificado en el relato infantil "*The Little Engine That Could*") de las variaciones en el desempeño de los estudiantes, una forma extrema del mito del desvanecimiento de la varianza al que me referí en el capítulo 6.[14] Es una enorme ingenuidad esperar que los estudiantes con retardo leve superen a todo el tercio inferior de los alumnos de los países con mejores resultados del mundo y esperar encima que esto suceda en 12 años. No dudo que esta exigencia esté motivada por la buena intención de obligar a las escuelas a prestar más atención al logro de los alumnos con discapacidades. Esto contrasta con los requisitos federales anteriores que se enfocaban en cuestiones de procedimiento, como las ubicaciones apropiadas en el aula para los alumnos. No obstante, como antiguo maestro de educación especial, considero que la condición extrema de esas exigencias es injusta con los maestros y cruel con los estudiantes, porque los obliga a presentar exámenes en que no pueden tener éxito y a recibir la etiqueta de fracaso incluso si su trabajo es bueno en relación con sus capacidades.

Ofreceré un ejemplo para concretar lo anterior. Durante el tiempo en que trabajé como maestro de educación especial, una de mis tareas consistía en enseñar lectura remedial. Sin lugar a dudas, esa fue la labor docente más difícil que haya realizado nunca y no creo que fuera muy bueno en ella. Tenía alumnos de quinto grado que leían al nivel de segundo; en términos muy aproximados, su tasa de adquisición de las habilidades de lectura equivalía posiblemente a la tercera parte de la tasa promedio. Suponga ahora que yo hubiera manejado las cosas para *duplicar* la tasa a la que obtenían esas habilidades, lo cual habría sido un éxito notable para mí y para mis alumnos. Los chicos habrían avanzado de estar muy rezagados del nivel de grado a un rezago modesto, pero según las nuevas reglas, tanto ellos como yo habríamos sido un fracaso.

Dos años después de la publicación de esas regulaciones, el Departamento de Educación de Estados Unidos publicó otra reseña de la reglamentación propuesta que disminuía un poco la severidad de esos requisitos, los cuales se concluyeron en abril de 2007.[15] Las regulaciones revisadas reconocían que existen otros estudiantes aparte del 1 por ciento cuyas discapacidades hacen poco práctico esperar que alcancen la competencia al nivel del grado incluso con la mejor enseñanza, y permitían a los estados aplicar "estándares de logro modificados" en la evaluación un máximo de un 2 por ciento adicional de los estudiantes. Esos estándares modificados reflejarían "una menor amplitud o profundidad del *contenido del nivel de grado*" (énfasis agregado).[16] La manera en que esto operaría en la práctica es un enigma. En las materias en que el currículo es acumulativo, como la lectura y matemáticas básicas, la mayor parte de los alumnos de bajo logro no podrían mantener el ritmo de sus coetáneos y a la postre terminarían por estudiar material de un grado inferior.

Sin embargo, los detalles de las regulaciones son mucho menos importantes que el dilema fundamental que ponen de relieve: la dificultad de decidir la mejor manera de examinar a los estudiantes de bajo logro. Por un lado, yo y muchos otros consideramos que el impulso actual para mejorar el desempeño de los estudiantes de bajos resultados es esencial y debería haberse adoptado hace mucho tiempo. Esto requiere de estándares más altos para esos alumnos. Al mismo tiempo, incluso si tuviésemos éxito en este aspecto, todavía tendríamos que enfrentar una distribución muy amplia del desempeño. El dilema es encontrar una manera de alcanzar la meta de encarar variaciones no deseadas en el desempeño al mismo tiempo que somos realistas acerca de las variaciones que persistirán.

Estudiantes con un nivel de competencia limitada del inglés

Los problemas que surgen al examinar a estudiantes con un nivel de competencia limitada del inglés son, en algunos sentidos, sorprendentemente similares a los que enfrentamos al examinar a los alumnos con discapacidades. Pero también hay varias diferencias importantes.

Mi propia experiencia de vivir en Israel con un dominio limitado del hebreo puede ilustrar lo anterior, aunque es necesario ir más allá de mis vergonzosas metidas de pata y mi discurso aburrido. Otra limitación de mi competencia, más pertinente para esta exposición, es que de haber presentado exámenes en hebreo mi desempeño habría sido muy pobre. Lo entendí a tiempo, porque estaba pensando en hacer el trabajo de posgrado en Israel y sabía muy bien que no tenía el dominio del idioma necesario para el estudio serio. Pero mis limitaciones con el idioma me resultaron todavía más claras hace poco cuando impartí una serie de seminarios en Jerusalén acerca de temas de evaluación y tuve ocasión de revisar varias formas del PET, la prueba de admisión a la universidad en hebreo que es análoga en muchos sentidos a la prueba SAT. De haber presentado en ese momento esa prueba sin adecuaciones, habría terminado con una puntuación terriblemente baja. Habrían sido muchos los reactivos que ni siquiera hubiera podido leer y muchos otros que sólo habría podido leer con mucho tiempo adicional.

Con estos antecedentes, supongamos que hubiera presentado la prueba PET y que luego hubiese solicitado el ingreso a una universidad israelita. ¿Qué habrían concluido sobre mí los funcionarios de la oficina de admisiones a partir de mi pésimo resultado?

Si hubiesen inferido que carecía de las habilidades matemáticas y de otras habilidades cognoscitivas necesarias para el estudio universitario, habrían estado equivocados. Si esta fuera la inferencia

proyectada, entonces la puntuación que yo hubiese obtenido, sin adecuaciones, habría estado muy sesgada. Suponga que los funcionarios de la oficina de admisiones quisieran una respuesta a una pregunta similar: si yo podría ser un estudiante competente en un programa universitario en hebreo de contar con más tiempo y estudio del idioma. También en este caso mi puntuación tendría un sesgo hacia abajo, lo que les daría una respuesta demasiado deprimente.

Suponga ahora que quisieran responder una tercera pregunta: si *en ese momento y con el nivel de competencia que tenía entonces* era probable que yo tuviera éxito al estudiar en una universidad con idioma hebreo. En ese caso, mi baja puntuación habría sido la correcta: yo hubiera sido un alumno realmente débil.

¿Cuál es la distinción entre las dos primeras preguntas, para las cuales mi puntuación habría sido sesgada, y la tercera, para la cual no lo habría sido? El problema es el mismo que surge cuando se examina a estudiantes con discapacidades: si el impedimento (sea que surja de una discapacidad o de una competencia limitada en el idioma del examen) es relevante para la pregunta que el resultado pretende responder. Aplicar la prueba PET sin adecuaciones para evaluar mis habilidades de matemáticas habría sido análogo a que yo aplicara el examen sin adecuaciones a mi alumna con acromatopsia. La razón para mi mal desempeño en la prueba PET y para el mal desempeño de esa chica en mi examen no habría sido pertinente para la inferencia que se basa en los resultados obtenidos en la prueba, y por lo tanto esas puntuaciones habrían estado sesgadas. Usar la prueba PET sin adecuaciones para evaluar si *en ese momento* yo tenía la capacidad para tener un buen desempeño en una universidad con idioma hebreo es más parecido a evaluar a un estudiante disléxico en lectura. Para responder esas preguntas, mi limitado dominio del hebreo y la dislexia del alumno son relevantes para lo que se trata de medir.

Suponga ahora que hubiera presentado la prueba PET en inglés (ellos tienen formas traducidas) o con algunas otras adecuaciones para compensar mi dominio limitado del hebreo. (No creo que cualquier otra adecuación hubiera sido suficiente, pero en aras de la discusión, suponga que sí.) Entonces, la puntuación que hubiera obtenido con la adecuación habría proporcionado una mejor respuesta a la primera pregunta: yo tenía las habilidades cognoscitivas necesarias para el estudio universitario, aunque mi hebreo fuese primitivo. Pero por eso mismo, la calificación obtenida con la adecuación habría generado una calificación con un sesgo *ascendente* para los propósitos de la tercera pregunta: es decir, habría hecho una predicción demasiado optimista respecto a lo bien que me habría ido en la universidad en ese año.

Este ejemplo aclara que existen varias semejanzas importantes en la evaluación de los alumnos con discapacidad y los que tienen una competencia limitada en el inglés. Una es que no siempre hay una manera "correcta" de evaluar a esos estudiantes. La mejor manera de evaluarlos (si deben usarse traducciones, si deben ofrecerse otras adecuaciones, etcétera) depende de la inferencia que se pretenda apoyar con los resultados de la prueba. Y tenemos que ser mucho más específicos de lo que acostumbramos acerca de las inferencias que se buscan. No basta referirse a "competencia en matemáticas" o "preparación para los estudios universitarios".

Además, al evaluar a ambos grupos enfrentamos impedimentos que son irrelevantes y relevantes para los constructos que intentamos medir. Cuando esas barreras son irrelevantes (como en el caso de mi alumna con acromatopsia) podemos tratar de compensarlas con adecuaciones. Cuando los impedimentos son relevantes para lo que tratamos de medir (como en el caso de la aplicación de la prueba de lectura a un estudiante disléxico) enfrentamos un problema lógico, no sólo técnico, y es poco probable que las adecuaciones lo resuelvan por completo.

Existen otras similitudes importantes entre esos dos grupos de estudiantes. Una es que la investigación que explora la mejor manera de examinarlos todavía es muy limitada. En años recientes, varias personas han estudiado tanto los efectos de las adecuaciones para los estudiantes con un dominio limitado del inglés como los aspectos del diseño de la prueba que podrían disminuir las barreras irrelevantes para el constructo que enfrentan. Por ejemplo, si se evitan complejidades lingüísticas innecesarias en las pruebas de matemáticas (como el uso de la voz pasiva, los tiempos verbales complejos, palabras que se usan muy poco o un exceso de expresiones idiomáticas) se podrían disminuir las dificultades que enfrentan los alumnos con dominio limitado del inglés. Sin embargo, esta investigación todavía es incipiente.

Una anécdota de uno de mis grupos pone de manifiesto el problema de las expresiones idiomáticas. Un año, más de la mitad de los alumnos de uno de mis cursos avanzados de metodología eran estudiantes extranjeros. Según cualquier definición razonable, todos hablaban inglés con soltura; después de todo, estaban realizando con éxito estudios de posgrado en Harvard sin apoyo especial. Un día, a la mitad de la sesión, vi que una estudiante de Chile se inclinaba y murmuraba algo a la chica que estaba a su lado, que era de Brasil pero también hablaba español. A su vez, ella se inclinó hacia la estudiante a su lado, que era de Venezuela, y susurró algo. La estudiante venezolana (quien, da la casualidad que desde entonces se ha dedicado al estudio de los efectos de la complejidad de los reactivos de pruebas de matemáticas en el desempeño de estudiantes con dominio limitado del inglés) se inclinó sobre dos hablantes nativos para cuchichearle algo a un estudiante mexicano. En ese punto decidí que era momento de detener la clase y le pregunté a la primera estudiante cuál era el problema. Ella me miró con gran desconcierto y dijo: "¿A qué se *refiere* cuando dice que algo es '*small potatoes*' (una cosa trivial)?" Ni siquiera me había dado

cuenta de haber usado ese modismo. Ningún autor competente de pruebas lo habría utilizado en una prueba de matemáticas, pero es común que los hablantes nativos usemos complejidades lingüísticas más sutiles sin siquiera percatarnos. Por ejemplo, enfrentamos el problema de la *polisemia*, el hecho de que muchas palabras tienen significados múltiples no relacionados. Para los hablantes nativos es sencillo hacer cambios entre ellos y darse cuenta, por ejemplo, que "recortar un precio" significa "reducirlo" y no recortarlo como se hace con unas tijeras. Para los hablantes de un segundo idioma esto resulta mucho más difícil, ya que a menudo sólo conocen el significado más común de una palabra.

Una última semejanza: los estudiantes con dominio limitado del inglés, igual que los estudiantes con discapacidades, son un grupo heterogéneo ya que hablan literalmente cientos de lenguas maternas de muchos grupos lingüísticos que en esencia son diferentes. Por ejemplo, aunque el alemán es un idioma con muchas más desinencias que el inglés, sus estructuras verbales son muy parecidas y me resulta sencillo aprenderlas. En contraste, el hebreo tiene menos tiempos verbales pero muchas otras "construcciones" verbales que no tienen paralelo en inglés; y el chino no usa tiempos verbales en absoluto. Es probable que las dificultades que encaran los estudiantes con dominio limitado del inglés cuando presentan exámenes en ese idioma difieran dependiendo de la estructura de su lengua materna. Por ejemplo, existen investigaciones que demuestran que sustituir palabras con raíces alemanas por palabras con raíces latinas facilita los reactivos para los hispanoparlantes nativos (un hallazgo no sorprendente) pero es improbable que eso fuese de mucha ayuda para un hablante nativo del coreano, que no comparte raíces con ningún idioma. Sin embargo, apenas se están iniciando los estudios que investigan los efectos de esas diferencias en el desempeño en las pruebas.

También hay diferencias importantes entre los dos grupos de estudiantes. Una es que el problema de las barreras relevantes al constructo para el desempeño en las pruebas afecta a todos los estudiantes con dominio limitado del inglés en muchas pruebas, pero sólo representa una seria dificultad para un subconjunto de los estudiantes con discapacidades. Una segunda diferencia es que, si bien muchas discapacidades son persistentes, las dificultades enfrentadas por algunos estudiantes con dominio limitado del inglés –aquellos que en verdad son aprendices de ese idioma– disminuirán con el paso del tiempo. Es posible que nunca desaparezcan del todo, y no sabemos mucho acerca de la rapidez con que algunas disminuyen, pero es claro que a la mayoría de los estudiantes que han estado en Estados Unidos por algún tiempo les va mucho mejor que cuando recién llegaron.

Enfrentamos obstáculos en verdad fabulosos al evaluar a los estudiantes con discapacidades y a los que tienen un dominio limitado del inglés. En algunos casos, los resultados de las pruebas sólo pueden apoyar conclusiones limitadas acerca del nivel de competencia –más limitadas de lo que nos gustaría– y en otros, por el momento simplemente no podemos obtener buenos resultados. Sin embargo, esas dificultades no son razón para tirar la toalla. Si usamos las pruebas con cuidado, estando al tanto de las limitaciones inherentes a la evaluación de esos grupos, podemos obtener información útil acerca del desempeño de los alumnos, incluso si esa información es algo más limitada y ambigua de lo que nos gustaría. De igual modo, si tenemos el cuidado suficiente, la evaluación puede fomentar beneficios educativos importantes para esos alumnos. Al mismo tiempo, ignorar o restar importancia a los problemas inevitables en la evaluación de esos estudiantes dará como resultado información engañosa y correremos el riesgo de dañar precisamente a los estudiantes que queremos ayudar. ■

El ABC
de la
evaluación educativa

Capítulo 13
Usos razonables de las pruebas

Ahora puede entender a la estudiante que mencioné en el capítulo 1, la que dijo que se sentía "terriblemente frustrada" por la pérdida constante, día a día, de las respuestas simples y directas en torno a la evaluación. Recordará lo que le respondí: que el propósito de aprender sobre esos temas, en principio desalentadores, era generar una mayor comprensión de la evaluación que les permitiera, a ella y a sus compañeros, hacer un uso más productivo de las pruebas. Sostuve que sin la comprensión de los principios y conceptos fundamentales de la evaluación no es posible dar sentido a la información proporcionada por las pruebas o llegar a soluciones inteligentes de los numerosos e intensos debates que hay acerca de la evaluación en nuestras escuelas.

Al llegar a este momento debe estar bastante claro el riesgo de entender mal las puntuaciones obtenidas en las pruebas, pero ¿cómo puede uno aplicar los principios descritos en este libro para hacer una buena interpretación de los resultados y tomar mejores decisiones sobre el uso de las pruebas? No existe una receta única, por supuesto, ya que las pruebas tienen muchos usos distintos y se emplean en muchos contextos diferentes. No obstante, podemos seguir algunas directrices generales.

Empecemos con lo que puede hacer uno de los usuarios finales de los resultados (un padre, un escritor, un educador, un contribuyente) para efectuar una interpretación sensata que le permita obtener información más útil y evitar los malos entendidos

graves. Una prueba, incluso una muy buena, siempre es sólo una prueba: una fuente valiosa de información, pero aun así una visión limitada y particular del desempeño del estudiante. Con eso en mente, ¿qué factores podrían amenazar la inferencia específica que necesita? ¿Cómo puede lidiar con ellos para obtener una conclusión confiable?

La lista de amenazas a las conclusiones basadas en los resultados obtenidos en la prueba (las amenazas a la validez) es larga. Algunas de las mayores son las siguientes:

- Para empezar, está el error de medición, el cual crea una banda de incertidumbre alrededor de la puntuación de cada estudiante.
- Cuando nos preocupan los agregados, como la calificación promedio o el porcentaje del nivel de competencia en una escuela, también está el error de muestreo, que ocasiona fluctuaciones sin sentido en los resultados de un grupo de estudiantes a otro y de un año al siguiente. Este es un problema particularmente grave para los grupos pequeños, por ejemplo, cuando seguimos la trayectoria del desempeño de escuelas pequeñas o, aun más problemático, el desempeño de grupos de estudiantes dentro de una escuela.
- Los resultados en que confiamos a veces son específicos a una determinada prueba. Las diferencias en la elección de contenido, en los métodos de calificación, en los formatos de los reactivos e incluso en los métodos matemáticos para el escalamiento de una prueba pueden producir diferencias en los patrones de los resultados.
- Las diferentes formas de reportar el desempeño no siempre pintan el mismo cuadro. Esto es preocupante a la luz de la dependencia actual en el reporte basado en estándares, que es una de las peores maneras de reportar el desempeño en las pruebas y a veces es sencillamente engañoso.

- El sesgo potencial siempre debe ser una preocupación, en especial cuando se evalúa a ciertos grupos de estudiantes, como los que padecen discapacidades o tienen un dominio limitado del inglés.
- La presión actual (generalizada e intensa) por elevar los resultados crea el potencial para que estos sean gravemente inflados.

¿Cómo puede uno evitar tropezar con una lista tan larga de problemas potenciales? Primero, siendo cuidadoso acerca de la inferencia que se extrae. Segundo, buscando información adicional. Vamos a considerar algunos casos específicos.

En el capítulo 5 revisé algunas comparaciones internacionales de los resultados obtenidos en las pruebas, las cuales se han convertido en una influencia sumamente poderosa en el debate público y en la política, no sólo en Estados Unidos sino también en muchas otras naciones del mundo. Lo que más parece interesar a la gente es la carrera de caballos, la clasificación de los países en términos del desempeño de sus estudiantes. Esas clasificaciones a menudo se expresan en términos de un "promedio internacional". Con mucha frecuencia las conclusiones basadas en esos datos son algo imprecisas y sólo hacen referencia en términos generales a la competencia o desempeño en un área temática completa, como las matemáticas. Veamos la siguiente descripción que hizo el *New York Times* de los resultados del Estudio Internacional de Tendencias en Matemáticas y Ciencia *(Trends in International Mathematics and Science Study,* TIMSS*)*: "Los estudiantes estadounidenses de octavo grado tuvieron un mejor desempeño en matemáticas y ciencia el año pasado que en 1999, pero todavía van a la zaga de sus pares en otros países industrializados".[1] O la siguiente descripción de una aplicación anterior del TIMSS (tomada del sitio web del Departamento de Educación de Estados Unidos): "En 1999, los alumnos estadounidenses de octavo grado superaron el promedio

internacional de 38 países en matemáticas y ciencia".[2] Advierta que esas descripciones se refieren sólo a la competencia en matemáticas, no a una mezcla específica de contenido o habilidades matemáticas. En esas afirmaciones no hay nada que implique, por ejemplo, que la representación del álgebra –que en esta evaluación era mucho mayor que en la otra evaluación internacional importante, la prueba PISA– fuera apropiada o indeseable. ¿Qué amenazas a esas inferencias son las más importantes?

Para empezar, la idea de un "promedio internacional" es inútil. El promedio puede variar mucho de un estudio a otro, dependiendo de la mezcla de naciones participantes. Como indiqué en el capítulo 5, Estados Unidos estaba por arriba de la media nacional en una parte del informe TIMSS y por debajo en otra parte del *mismo* informe, debido a que cada sección utilizó una muestra diferente de naciones. Por ende, el primer paso consiste en ignorar las afirmaciones acerca del desempeño nacional en relación con el "promedio internacional" y en su lugar concentrarse en comparaciones específicas que pueden ser más informativas. Por ejemplo, puede ser útil contrastar el desempeño de los estudiantes estadounidenses con el de los alumnos de países de alto desempeño del Oriente Asiático, o compararlos con los estudiantes de países más similares, como Inglaterra y Australia. Sin embargo, ese es sólo un primer paso. Una vez que se ha dado ¿cómo podemos ponernos en un terreno seguro?

En el caso de las comparaciones internacionales, el error de muestreo es una de las amenazas más sencillas de enfrentar. Los estudios usados para comparar el desempeño de los estudiantes entre países (de manera más notable los estudios TIMSS y PISA) abordan con cuidado el error de muestreo. Si sólo queremos comparar promedios, podemos dejar de lado el error de muestreo siempre que las diferencias entre las medias sean lo bastante grandes para ser estadísticamente significativas, y los informes nos indiquen cuáles lo son.

Los resultados que son específicos a la prueba utilizada representan una preocupación mayor. En el capítulo 5 demostré que las pruebas PISA y TIMSS clasifican a los países de manera muy diferente y que dichas clasificaciones pueden ser modificadas, en un grado modesto, incluso por los cambios en el énfasis dado a las áreas de contenido incluidas en cualquiera de las pruebas. Por lo tanto es prudente ignorar las diferencias pequeñas, que es menos probable que sean consistentes entre diferentes pruebas o incluso entre diferentes ponderaciones de las áreas de contenido de una sola prueba. El hecho de que una diferencia sea estadísticamente significativa no es protección suficiente toda vez que el cálculo de la significancia estadística no toma en consideración las variaciones en los resultados entre pruebas.* Sería mucho más seguro concluir que Estados Unidos va a la zaga de Japón que inferir que nuestros estudiantes se desempeñan mejor que los de Eslovenia. La diferencia entre Estados Unidos y Eslovenia fue estadísticamente significativa en el estudio TIMSS realizado en 2003, pero su tamaño fue una sexta parte del tamaño de la brecha entre Estados Unidos y Japón.

Eso sería más seguro, pero no del todo. En ocasiones, resulta que incluso las grandes diferencias no son constantes de una prueba a otra. En la evaluación TIMSS de matemáticas de octavo grado realizada en 2003, el resultado de Noruega estuvo muy por debajo del de Estados Unidos. Según cualquier estándar razonable, la

* Una nota para el lector con orientación técnica: en la mayor parte de los reportes estadísticos, la significancia estadística sólo toma en consideración el error de muestreo. En algunas de las evaluaciones más complejas (de las que fue pionera la Evaluación Nacional del Progreso Educativo), el cálculo de la significancia estadística toma en cuenta tanto el error de muestreo como el de medición. Sin embargo, el error de medición surge de las fluctuaciones en el desempeño en pruebas alternativas del *mismo* diseño, como las formas alternativas de la prueba SAT. Lo que aquí nos preocupa es otra cosa: las variaciones en el desempeño en pruebas de diseños *diferentes*, como las diferencias entre las pruebas TIMSS y PISA.

diferencia era muy grande, equivalía casi a dos terceras partes de la brecha entre Estados Unidos y Japón. Sin embargo, en la evaluación PISA realizada el mismo año, Noruega superó a Estados Unidos, no por mucho, pero sí lo suficiente para que fuese estadísticamente significativa. Es posible que varios factores contribuyeran a esta sorprendente discrepancia en los resultados, como las diferencias en el contenido de la prueba, el nivel de edad (los estudiantes de la prueba PISA eran casi dos años mayores) y la construcción de la muestra (TIMSS hizo el muestreo por grado escolar mientras que PISA hizo el muestreo por edad). A pesar de eso, las personas que consideran que cualquiera de esas evaluaciones es la respuesta definitiva, el resumen "correcto" del logro relativo de las naciones en "matemáticas" (sin ninguna clasificación acerca de la mezcla de contenido implícita en las "matemáticas") estarían sobre un terreno muy superficial.

Y esto indica una de las mejores formas de evitar el mal uso de los datos de las pruebas: nunca considere que una sola prueba proporciona la respuesta "correcta" y fidedigna. Cuando sea posible, use más de una fuente de información acerca del logro (resultados de otras pruebas o información de otras fuentes). Con datos de varias fuentes (PISA, varias repeticiones de la prueba TIMSS y algunos estudios internacionales anteriores) podemos ver que no hay mucho lugar a dudas: Estados Unidos siempre obtiene resultados muy por debajo de Japón, aunque no siempre obtiene resultados por arriba de Noruega.

Cuando no se disponga de datos adicionales (una situación que, por desgracia, es demasiado común en la actualidad) tiene que retroceder al primer enfoque: ser muy cuidadoso acerca de su conclusión. Proteja sus apuestas. Considere la información de la única prueba como una instantánea del desempeño, la cual por fuerza es incompleta y tal vez sea un tanto diferente de la que habría obtenido de contar con otra medida que también sea razonable. Esto es

cierto incluso cuando su prueba es muy buena. Una forma más precisa de expresar la conclusión ofrecida por el *New York Times* habría sido: "En 1999, los estudiantes estadounidenses de octavo grado obtuvieron mejores puntuaciones en matemáticas, según la medición que hace la prueba TIMSS de esta materia". Esto no sería de ayuda para la mayor parte de los lectores del *Times* (no tendrían idea de lo que significa la advertencia), pero es la salvedad que usted debería tener en mente.

Con o sin datos adicionales, para proteger sus apuestas también debería evitar la precisión espuria. Considere la brecha entre Estados Unidos y Japón. Siempre aparece y siempre es grande, pero no siempre es exactamente del mismo tamaño. Es más seguro concluir que "en matemáticas de octavo grado, el resultado promedio de Japón por lo general está muy por arriba que el de Estados Unidos, casi una desviación estándar completa". No es seguro concluir que "la diferencia promedio en matemáticas de octavo grado entre Estados Unidos y Japón es de 0.83 desviaciones estándar" (el resultado de la prueba TIMSS en 2003). La última estimación más específica sólo se justifica si quiere cargarse con una inferencia más delimitada y menos interesante: "Se estimó que la diferencia promedio en matemáticas de octavo grado entre Estados Unidos y Japón, *específicamente en la prueba TIMSS aplicada en el 2003*, fue de 0.83 desviaciones estándar, con un margen de error de..."

Considere un ejemplo que es más polémico que la interpretación de las comparaciones internacionales: dar sentido a los resultados de las pruebas exigidas por el estado que ahora se utilizan para hacer responsables a los maestros (y a menudo a los estudiantes). Se han generalizado los grandes incrementos en los resultados y casi siempre se presentan como indicadores claros y veraces de que el logro de los estudiantes ha mejorado a ritmo acelerado. ¿Está esa inferencia garantizada?

El error de muestreo es una preocupación mayor en este caso que en la interpretación de los estudios internacionales. Los resultados estatales reflejan el desempeño de muchos estudiantes, por lo que el error de muestreo no es una amenaza grande para ellos. Sin embargo, muchas de la inferencias importantes corresponden a grupos más pequeños: estudiantes de un grado en una sola escuela o, más extremo, los subgrupos de estudiantes cuyo desempeño debe reportarse por separado según las exigencias de la ley NCLB, como el caso de los estudiantes con dominio limitado del inglés de un determinado grado de una sola escuela. Es común que esos resultados se basen en un pequeño número de observaciones. La figura 7.3 del capítulo 7 muestra que incluso si los subgrupos no se reportan por separado, el desempeño de las escuelas pequeñas es sumamente inestable y el de los subgrupos por lo general es mucho más errático. De modo que debería prestar poca atención a los cambios que se dan de un año a otro en los grupos pequeños, incluso para escuelas completas, y en su lugar examinar las tendencias a lo largo de varios años. Por desgracia, es poco probable que vea datos presentados de esta manera en los trabajos o incluso en los sitios web de los departamentos de educación de la mayoría de los estados. Por lo regular la carga recaerá en usted.

Sin embargo, con todo lo importante que es el error de muestreo, no es el mayor problema. La amenaza más grave a la validez de las inferencias hechas a partir de las pruebas de alto impacto es el riesgo de inflación de los resultados. Admito que no son muchas las investigaciones que abordan este problema, pero nuestra investigación demuestra que los resultados en las pruebas de alto impacto pueden inflarse con rapidez, a menudo en un monto muy grande. El mejor consejo basado en la investigación disponible es que no deben tomar al pie de la letra las mejoras en los resultados obtenidos en pruebas de alto impacto. Pueden ser un indicador de mejoras reales de los estudiantes, pero no debe confiarse en ellas

hasta que sean confirmadas por otros datos, los cuales pueden provenir de otras pruebas de logro (por ejemplo, de la NAEP o de una segunda prueba de bajo impacto aplicada por el distrito o por el estado) o de otras fuentes, como las pruebas de admisión a la universidad y las tasas de asignación a cursos remediales a nivel posterior a la secundaria y media superior. Muestre especial cautela ante las mejoras enormes y vertiginosas, las cuales son cada vez más comunes.

Este es sólo otro caso en que se confía en más de una sola medida de logro, pero con un giro. En el caso de las evaluaciones internacionales debe tenerse cuidado acerca de pruebas alternativas que proporcionan visiones algo distintas del desempeño en cualquier momento. En el caso de las pruebas estatales, el análogo del riesgo de las diferencias entre las pruebas PISA y TIMSS es el riesgo de que, si un estado contratara a dos vendedores para escribir los nuevos exámenes, estos podrían clasificar de manera distinta a los estudiantes, las escuelas o los distritos. El problema de la inflación de los resultados es diferente. Incluso si dos pruebas en un *inicio* arrojan resultados similares, tal vez dejen de hacerlo una vez que se sienten los efectos del alto impacto.

Esto es precisamente lo que sucedió en Kentucky en la década de los noventa. Cuando el estado introdujo su prueba de alto impacto para bachillerato, el Departamento de Educación demostró que existía un grado de consistencia moderadamente alto entre los resultados de esa prueba y los de la prueba ACT, que es la que más se utiliza en Kentucky para admisión a la universidad, y aseguró a los padres que "no es demasiado arrogante suponer que el mayor aprendizaje que da lugar a la mejora en una prueba es probable que conduzca a la mejora en la otra".[3] Tal vez no sea abiertamente presuntuoso, pero sí claramente optimista. En realidad, no se observó consistencia en el *cambio a lo largo del tiempo* mostrado por las dos pruebas; los resultados de matemáticas en la prueba

estatal se elevaron con rapidez, mientras que los resultados en la prueba de matemáticas de la ACT no mejoraron en absoluto.

Digamos que usted está interesado en saber qué tanto ha mejorado en total el logro de los estudiantes de su estado, o de manera todavía más específica, si se ha hecho algún progreso para disminuir la brecha del desempeño entre grupos raciales y étnicos. Busca los datos (o los encuentra en el periódico de la mañana) y descubre que todo se presenta en términos del porcentaje de estudiantes que alcanzan o superan algún estándar de desempeño, con más probabilidad del estándar de "competente". En el capítulo 8 esbocé una serie de razones de por qué esta es una manera inadecuada de reportar el desempeño: oculta mucha información, exagera la importancia de otra y distorsiona las comparaciones de las tendencias en el desempeño entre los grupos de estudiantes con puntuaciones más altas y más bajas. Por el momento dejemos de lado el problema anterior, que es el hecho de que los porcentajes que reciben la etiqueta de competente pueden ser muy exagerados. ¿Qué puede usted hacer para evitar los problemas inherentes al reporte basado en estándares?

Una opción es buscar una forma alternativa de reporte, que sea más informativa y menos propensa a distorsionar las tendencias que le interesan. Muchos estados disponen de escalas más razonables pero no las reportan (o les dan menos énfasis) debido a los requisitos de la ley NCLB y al entusiasmo generalizado, aunque mal dirigido, por el reporte basado en estándares. De hecho, muchos estados necesitan esas mejores escalas; son las que sus psicómetras utilizan para ubicar los estándares en niveles de dificultad comparables de un año al siguiente, a medida que cambian las formas específicas de la prueba. De modo que si busca un poco tal vez pueda obtener una escala de resultados que le permita comprender mejor las tendencias entre grupos. Y si usted es ambicioso y está dispuesto a hacer algunos cálculos a mano, quizá pueda obtener

incluso las desviaciones estándar de esas escalas, lo que le permitiría poner todas las diferencias y cambios que le interesan en fracciones de una desviación estándar, lo que las haría aproximadamente equivalentes con todo tipo de otros datos. El departamento estatal o local de educación puede incluso tener cálculos de ese tipo o estar dispuesto a proporcionarlos.

Pero digamos que no hay escalas de resultados disponibles o por lo menos ninguna que alguien quiera proporcionarle. ¿Qué hacer entonces? Sin algunas técnicas matemáticas que son demasiado complejas para presentarlas aquí, no hay nada que pueda hacer para reparar las distorsiones que puede crear el reporte basado en estándares. No obstante, puede evitar los otros riesgos importantes que genera esa forma de reporte: interpretar los estándares de modo que signifiquen más de lo que en realidad significan.

Tenga en mente lo que son (y lo que no son) los estándares de desempeño. Son sólo puntos de corte sobre un continuo de desempeño. Es común que los periódicos y las publicaciones de los departamentos de educación presenten los hallazgos como la descripción de grupos cualitativamente distintos, como el de los estudiantes que son "competentes" y el de los alumnos que no lo son. Sin embargo, mientras las diferencias entre los estudiantes apenas por arriba y apenas por debajo de un estándar son triviales, pueden ser enormes entre los estudiantes que caen entre dos de los estándares a quienes se asigna por ende la misma etiqueta. Más aún, el proceso de establecimiento de estándares, aunque misterioso y aparentemente "científico", no es una forma de revelar alguna verdad subyacente acerca de categorías del logro de los estudiantes. Los métodos usados son sólo una forma muy complicada de utilizar el juicio para decidir qué puntuación es lo bastante alta para merecer la etiqueta de "competente".

Para ayudarse a evitar una interpretación exagerada de los estándares de desempeño, podría cambiarles la etiqueta. En lugar

de llamarlos debajo de lo básico, básico, competente y avanzado –etiquetas que acarrean una carga indeseable–, intente considerarlos como cuatro niveles de desempeño meramente arbitrarios, digamos, nivel 1, nivel 2, nivel 3 y nivel 4. Los defensores del reporte basado en estándares podrían decir que esta sugerencia es exagerada y que los estándares de alguna manera están ligados a las descripciones de lo que los niños pueden hacer en realidad. Existe algo de verdad en ese reclamo, pero el hecho incómodo es que los diversos métodos usados para establecer los estándares de desempeño pueden ser sorprendentemente incongruentes. Algunos de los estudiantes que son "competentes" cuando se utiliza un método no lo serán cuando se prueba con otro, incluso si las definiciones de competente son idénticas. Además, la mayor parte de los métodos empleados están muy lejos de examinar el verdadero trabajo de los estudiantes reales. No hay razón para esperar que si usted y sus amigos ponen en fila a 100 estudiantes, los ordenan del desempeño más bajo al más alto, y examinan su trabajo, terminarán por colocar el corte de "competente" en un lugar cercano al punto en que lo colocó el departamento de educación de su estado utilizando el método del marcador, el método Angoff modificado o cualquier otro. Me parece más probable que usted se confunda si acepta las descripciones de los estándares tal como se ofrecen que si trata a los estándares como clasificaciones arbitrarias.

Por último, ¿qué puede hacer en relación con los intentos generalizados de utilizar los resultados de las pruebas como un indicador simple de la eficacia o calidad de la escuela? Intente el método de Nancy Reagan: "Sólo di no". Hay tres razones distintas por las que los resultados obtenidos en una prueba, tomados por sí mismos, son insuficientes para informar sobre cuáles son buenas y cuáles malas. La primera es que incluso una muy buena prueba de logro es necesariamente incompleta y dejará sin medir muchos aspectos de la calidad de la escuela. Algunos defensores a ultranza

de las pruebas de alto impacto menosprecian este argumento por considerarlo "contrario a la evaluación por la vía de pruebas", pero es una simple declaración de hechos que durante generaciones ha sido reconocida dentro de la profesión de la evaluación.

La segunda razón para no suponer que las puntuaciones altas necesariamente identifican a las mejores escuelas es el hecho de que, en el clima actual, pueden existir diferencias muy grandes entre escuelas en la cantidad de inflación de las puntuaciones. Algunas escuelas toman más atajos que otras en la carrera para elevar los resultados, y los trabajos están llenos de historias inverosímiles acerca de escuelas que lograron enormes progresos en un tiempo muy corto.

La tercera razón, y quizá la más importante, por la cual los resultados de las pruebas no pueden decirle si una escuela es buena o mala es que las escuelas no son la única influencia sobre los resultados. Otros factores, como el nivel y las expectativas educativas de los padres, tienen gran impacto sobre el desempeño de los estudiantes. Separar el impacto de la calidad de la escuela de los poderosos efectos de las muchas influencias extraescolares sobre el logro es una tarea muy difícil que no puede llevarse a cabo con los datos que suelen estar disponibles para los sistemas escolares. El resultado es que no se puede asumir de manera segura que las escuelas con los mayores aumentos en los resultados en verdad están mejorando con mayor rapidez ni que las que obtienen las puntuaciones más altas son las mejores.

¿Cómo debería entonces usar los resultados para ayudarse a evaluar a una escuela? Empiece por recordar que los resultados describen algo de lo que los estudiantes pueden hacer, pero no describen todo lo que pueden hacer y *no explican por qué pueden o no pueden hacerlo*. Utilice los resultados como punto de partida y busque otra evidencia sobre la calidad de la escuela; idealmente no deberían ser sólo otros aspectos del logro de los estudiantes sino

también de la calidad de la instrucción y de otras actividades dentro de la escuela. E investigue por sí mismo. Si los estudiantes obtienen buenos resultados en las pruebas de matemáticas pero parecen aburrirse hasta las lágrimas en las clases de esa materia, tome sus altas puntuaciones con escepticismo, porque una aversión a las matemáticas les costará más tarde en la vida, incluso si sus puntuaciones en octavo grado eran buenas.

La lista es sólo ilustrativa. Tal vez usted quiera usar los resultados de las pruebas con un propósito totalmente diferente para el cual los riesgos más importantes de la mala interpretación son un tanto distintos de los anteriores. Por ejemplo, muchos maestros quieren obtener información diagnóstica acerca de las fortalezas y debilidades relativas del desempeño de sus estudiantes. (¿Tengo más éxito para enseñarles a calcular que a aplicar las matemáticas en la resolución de problemas?) Cuando yo estaba en la escuela, una función importante de la evaluación era proporcionar información de este tipo, y por obvias razones muchos maestros todavía la desean, incluso si el mundo político se interesa mucho más en un simple juicio sumario del nivel de desempeño general de los estudiantes. Este uso de los resultados plantea un problema diferente: ¿son las diferencias en el desempeño en diferentes partes de la prueba lo bastante confiables para que puedan usarse como base para modificar la instrucción?

La tarea de hacer una buena interpretación de los resultados se parece un poco a tocar jazz. La clave para la improvisación es conocer los cambios de acordes. Los principios explicados en los capítulos anteriores son los cambios de acordes y usted necesita hacer la improvisación, pensar con cuidado en las amenazas que se ciernen sobre las inferencias que quiere hacer y en la mejor manera de mantenerse en un terreno seguro.

Esos principios también tienen implicaciones para quienes controlan o querrían influir en el diseño de los programas de

evaluación de sus escuelas. El primer consejo que daría a quienes toman las decisiones acerca de la evaluación es evitar las expectativas poco realistas. Esto puede denominarse «el principio de los Rolling Stones»: "No puedes obtener siempre lo que deseas... y si alguna vez lo intentas, recibirás lo que necesitas". Las expectativas poco realistas acerca de la evaluación se encuentran en todos lados. Parecen estar basadas en una perspectiva incongruente, incluso paradójica, de las complejidades de la medición y del consejo ofrecido por personas como yo. Por un lado, por lo general se pasan por alto las complejidades de la evaluación y a menudo se ridiculizan las complicaciones planteadas por los expertos como algo demasiado misterioso para que importe. Pero, por otro lado, parece existir una fe generalizada en los prodigios de la psicometría, la creencia tácita de que sin importar lo que los políticos y los educadores quieran que haga una prueba, de alguna manera podemos averiguar cómo hacer que funcione.

Una expectativa irracional muy generalizada es que una prueba creada para un propósito funcionará bien para muchos otros. Pero una sola prueba no puede servir a todos los amos. Recuerde que una prueba es una pequeña muestra de un dominio grande. La gente de mi campo que carga la tarea de crear una de esas muestras debe diseñar la prueba que mejor cumpla las metas más importantes de quienes la solicitan. El diseño y la elaboración de la prueba implican una larga serie de compensaciones y compromisos. De manera invariable, servir bien a un amo significa servir mal a otros. Por ejemplo, una prueba optimizada para proporcionar información sobre grupos no es la mejor para proporcionar puntuaciones de estudiantes individuales. Esa es la razón por la cual la Evaluación Nacional del Progreso Educativo no puede proporcionar resultados de individuos. Una prueba compuesta por un pequeño número de tareas grandes y complejas, en un esfuerzo por evaluar la competencia de los estudiantes para resolver problemas difíciles,

será inapropiada para identificar qué habilidades específicas han logrado dominar o no los estudiantes. Podría dar muchos otros ejemplos.

Quienes toman las decisiones deben determinar qué metas son más importantes para una prueba y luego aceptar el hecho de que ese resultado tendrá el costo de otras metas no incluidas. Este consejo poco grato suele ser ignorado y las consecuencias pueden ser importantes. Considere el requisito de la ley NCLB de que casi todos los estudiantes, sin importar su nivel de desempeño, sean evaluados con las mismas pruebas. Cualesquiera que sean las virtudes políticas de este requisito, es una mala medición: sabemos que un costo de diseñar una prueba que hace un buen trabajo al medir el desempeño de los estudiantes de alto logro es que hará un mal trabajo en el caso de los alumnos de bajo logro, y viceversa. Para obtener información válida y confiable acerca de lo que los alumnos están aprendiendo, es necesario que enfoquemos las pruebas en sus niveles de desempeño y en el contenido que estudian en la actualidad.

Otros ejemplos de expectativas poco realistas son las especificadas en leyes y regulaciones para la evaluación de estudiantes con necesidades especiales –los que presentan discapacidades y quienes tienen un dominio limitado del inglés. Esas regulaciones sostienen que debemos proporcionar a dichos estudiantes "adecuaciones apropiadas" y a menudo se espera que, con esas adecuaciones, los resultados de esos estudiantes tengan un significado comparable a los obtenidos por otros alumnos. Pero existen límites a nuestra capacidad para hacer que las puntuaciones obtenidas por algunos estudiantes con necesidades especiales sean en verdad comparables. En parte esto refleja la escasez de investigación y desarrollo, pero en ocasiones los problemas son lógicos, no técnicos. En esos casos, existen limitaciones en las inferencias que podemos sacar de manera segura acerca de los estudiantes, con o

sin adecuaciones para la prueba. Esto no es una razón para excluir a esos estudiantes de la evaluación, y ciertamente no es un argumento en contra de mayor investigación que nos ayude a evaluarlos mejor. Sin embargo, el beneficio para los estudiantes y sus maestros sería mayor si reconociéramos con franqueza las limitaciones de lo que podemos hacer, interpretáramos apropiadamente los resultados de los estudiantes y adaptáramos con base en ello nuestras respuestas educativas.

En algunos casos los objetivos de desempeño usados en los programas de evaluación de gran escala también son poco realistas. La idea de que podemos averiguar lo que los estudiantes "competentes" deben ser capaces de hacer y luego exigir a las escuelas que los ayuden a lograrlo resulta atractiva, pero como han demostrado los capítulos previos, la manera actual en que esto se realiza puede dar lugar a una débil estructura, como una casa construida de naipes. Si vamos a seguir usando las pruebas para establecer metas de desempeño para los maestros y las escuelas necesitamos encontrar mejores formas de hacerlo, aproximaciones que reflejen expectativas de progreso realistas y prácticas. Por ello tenemos que utilizar datos empíricos para establecer objetivos, no ideales vagos y generales. Esos datos pueden incluir información histórica acerca de las tasas de cambio, evaluaciones de programas ejemplares o detalles acerca de escuelas ejemplares. Al establecer las metas también es necesario reconocer que las grandes variaciones en el desempeño son una realidad humana, algo con lo que nuestro sistema educativo tendría que lidiar incluso si tuviéramos la voluntad política y los medios para reducir las flagrantes inequidades sociales que asolan a nuestro sistema educativo y nuestra sociedad.

Un consejo relacionado: así como recomiendo a los usuarios de las pruebas no confiar en estándares de desempeño para interpretar el nivel de competencia de los estudiantes, exhorto a quienes

tienen el control de los programas de evaluación a no obligar a los usuarios a hacerlo. El "porcentaje por arriba de competente" es un número arbitrario.

Oculta una gran cantidad de información y las tendencias en este porcentaje pueden ser gravemente engañosas. De ser importante establecer algunos estándares para reflejar expectativas, hágalo, pero no prive a los usuarios de los datos de otras formas de reporte que son más informativas y menos propensas a distorsiones. Reporte los estándares junto a datos en formas más útiles, como escalas de puntuaciones y percentiles.

Un último –y políticamente desagradable– consejo: es necesario que seamos más realistas en relación con el uso de las pruebas como parte de los sistemas de rendición de cuentas en la educación. Es poco probable que los sistemas que sólo presionan a los profesores para que aumenten los resultados en una prueba (o en un conjunto de pruebas de unas cuantas materias) funcionen como se anuncia, sobre todo si los incrementos exigidos son grandes e inexorables. Más bien es probable que produzcan una considerable inflación de los resultados y diversos cambios indeseables en la enseñanza, como un enfoque excesivo en antiguos exámenes, una restricción inapropiada de la instrucción y una dependencia de la enseñanza de trucos para presentar exámenes.

Apoyo con firmeza la meta de una mejor rendición de cuentas en la educación pública. Entendí que era necesaria cuando fui maestro de primaria y secundaria hace muchos años, y me resultó evidente como padre de dos niños en la escuela. En más de un cuarto de siglo de investigación educativa nada me ha hecho cambiar de opinión a este respecto. Y parece claro que el logro de los estudiantes debe ser una de las cosas más importantes de las que debe responsabilizarse a los educadores y a los sistemas escolares. Sin embargo, necesitamos un sistema de rendición de cuentas que sea eficaz, que maximice las ganancias reales y minimice las mejoras

falsas y otros efectos colaterales negativos. Necesitamos un sistema similar al de la FDA para la aprobación de fármacos: que sean eficaces y seguros. Todo lo que hemos visto hasta ahora nos dice que los sistemas actuales de rendición de cuentas basados en exámenes no satisfacen este estándar.

Aunque en el mundo actual de la política educativa se considera que esta advertencia sobre la rendición de cuentas basada en exámenes es extrema, es todo menos eso. La teoría económica predice tanto el engaño del sistema que hemos visto en numerosos estudios como el sesgo resultante en la medida empleada para la rendición de cuentas (en este caso, la inflación de resultados o calificaciones). Hemos visto que esos problemas surgen en muchos otros campos, áreas tan diversas como el cuidado de la salud, la regulación ambiental, los programas de capacitación para el trabajo y las estadísticas de delitos. De hecho, como se mencionó en el capítulo 10, la distorsión que surge cuando se utiliza una medida para la rendición de cuentas es tan común que ha llegado a conocerse como la ley de Campbell, en honor a un experto en programas de evaluación que escribió sobre el tema hace más de tres décadas.[4] No hay motivo para esperar que la rendición de cuentas educativa basada en exámenes sea una excepción de la ley de Campbell y la evidencia empírica indica que no lo es.

Es común que los defensores de los sistemas actuales de rendición de cuentas basada en exámenes repliquen con el argumento "¿Y qué si las mejoras son distorsionadas? Lo que importa es que los estudiantes aprendan más y si lo logramos, podemos vivir con alguna distorsión". Hipotéticamente así es, podríamos vivir con ella si supiéramos que en efecto los estudiantes están aprendiendo más y con distorsiones lo bastante pequeñas como para no inducir a la gente a errores graves que la hagan tomar decisiones incorrectas. Pero en realidad no sabemos cuánto ha cambiado el aprendizaje real de los estudiantes (o incluso si ha cambiado) como resultado

de esos programas. Dado que muchas personas consideran que los sistemas de rendición de cuentas son autoevaluativos (asumen que si las puntuaciones están aumentando podemos confiar en que los niños están aprendiendo más), hay una perturbadora carencia de buenas evaluaciones de esos sistemas, incluso después de más de tres décadas de evaluaciones de alto impacto. Como advertí en el capítulo 10, dicha evidencia como tal no deja cabida para el optimismo. Lo que sabemos es que la inflación de resultados puede ser enorme, lo suficiente para engañar gravemente a la gente. Más aún, por lo regular no podemos distinguir entre las mejoras reales y las falsas, por lo que no sabemos a qué escuelas recompensar, castigar o emular. Esto permite que los adultos que participan en el juego se declaren exitosos y que quienes se quedan rezagados sean sólo los estudiantes.

En general, la evaluación educativa se parece mucho a un medicamento potente. Si se utiliza con cuidado puede ser sumamente informativa y una herramienta muy poderosa para mejorar la educación. Si se usa de manera indiscriminada plantea el riesgo de varios efectos colaterales graves. Sin embargo, a diferencia de los medicamentos potentes, las pruebas se usan con poca supervisión independiente. Que el comprador tenga cuidado. ■

Notas

Capítulo 1. Si sólo fuera tan simple

1. J. Hassell, "Bush Hints at Compromise on Standardize Test Plan", *Seattle Times*, 15 de marzo, 2001.
2. Stephen P. Klein *et al.*, *What Do Test Score in Texas Tell Us?* Documento número IP-202 (Santa Monica, CA: Rand, 2002).

Capítulo 2. ¿Qué es una prueba?

1. F. J. Fowler, "How Unclear Terms Affect Survey Data", *Public Opinion Quarterly* 56, no. 2 (1992): 218-231.
2. H. J. Parry and H. M. Crossley, "Validity of Responses to Survey Questions", *Public Opinion Quarterly* 14, no. 1 (1950): 16-80.
3. Andrew Biemiller, "Oral Comprehension Sets the Ceiling on Reading Comprehension", *American Educator* 27 (Primavera 2001): 23.
4. Massachusetts Department of Education, *2005 MCAS Technical Report* (Malden, MA: Massachusetts Department of Education, 2006).
5. "Test Scores Move Little in Math, Reading: Improvement Appears Slight since No Child Left Behind", *Washington Post*, 20 de octubre, 2005, p. A03.

Capítulo 3. Lo que medimos: ¿Qué tan buena es la muestra?

1. E. F. Lindquist, "Preliminary Considerations in Objective Test Construction", en E. F. Lindquist, ed., *Educational Measurement* (Washington DC: American Council on Education, 1951).
2. H. D. Hoover *et al.*, *Iowa Test of Basic Skills, Interpretive Guide for School Administrators* (Chicago: Riverside Publishing, 2003).
3. Lindquist, "Preliminary Considerations in Objective Test Construction", 142.
4. Hoover *et al.*, *Iowa Test of Basic Skills, Interpretive Guide*.

Capítulo 4. La evolución de la evaluación en Estados Unidos

1. H. D. Hoover *et al.*, *The Iowa Test of Basic Skills, Interpretive Guide for School Administrators, Formas K and L, Niveles 5-14* (Chicago: Riverside Publishing, 1994).
2. Para las clasificaciones completas, junto con una tabla mostrando qué diferencias son estadísticamente significativas y por lo tanto dignas de confianza, vea: V. S. Mullis *et al.*, *TIMSS 2003 International Mathematics Report* (Chestnut Hill, MA: International Study Center, Boston College, 2004), 34, 38. Disponible en: *http://timss.bc.edu/timss2003i/intl_reports.html*.

3. "Standardized Test: What Are Trying to Measure?" *Talk of the Nation*, National Public Radio, 21 de marzo, 2002. *www.nrp.org/templates/story/story.php?storyId=1140228* (Mayo 24, 2006).
4. Para ejemplo, vea, Mullis *et al.*, *TIMSS 2003 International Mathematics Report*, exhibe D. 2, 412.
5. El impacto del NAEP y el Title I se evidenció hace aproximadamente dos décadas; vea P. W. Airasain, "State Mandated Testing and Educational Reform: Context and Consequences", *American Journal of Education* 95 (1987): 393-412, y E. Roeber, "A History of Large-scale Testing Activities at the State Level", documento presentado en the Indiana Governor's Symposium on ISTEP, Madison IN, febrero, 1988.
6. Vea R. M. Jaeger, "The Final Hurdle: Minimum Competency Achievement Testing," en G. R. Austin and H. Garber, eds., *The Rise and Fall of National Test Scores* (New York: Academic Press, 1982).
7. National Commission on Excellence in Education, *A Nation at Risk* (Washington, DC: U.S. Department of Education, 1983).
8. J. J. Cannell, "Nationally Normed Elementary Achievement Testing in America's Public Schools: How all 50 States Are above the National Average," *Educational Measurement: Issues and Practice* 7, no. 2 (1988): 5-9.
9. Individuals with Disabilities Education Act Amendments of 1997, 20 U.S.C.§1412(a)(17). No Child Left Behind Act of 2001, 20 U.S.C. 6311 *et seq*.
10. Por ejemplo, vea, R. L. Linn, "Assessments and Accountability", *Educational Researcher* 29, no. 2 (2000): 4-16.

Capítulo 5. ¿Qué nos dicen las calificaciones
de las pruebas sobre los niños estadounidenses?

1. National Commission of Excellence in Education, *A Nation at Risk* (Washington, DC: U.S. Department of Education, abril, 1983). Disponible en *www.ed.gov/pubs/NatAtRisk/index.html* (Recuperado el 8/9/05).
2. Paul E. Peterson, "A Ticket to Nowhere", *Education Next* (Primavera, 2003):39-46. Las citas son de las pp. 39-40.
3. National Center of Education Statistics, *The Condition of Education*, 2006 (Washington , DC: U.S. Department of Education, 2006), Apéndice 1, Tabla 7-1.
4. National Center of Education Statistics, *NAEP 2004 Trends in Academic Progress: Three Decades of Performance in Reading and Mathematics, Findings in Brief* (Washington, DC: U.S. Department of Education, 2005), 8.
5. Archived information, *Policy Brief: What the TIMSS Means for Systemic School Improvement*, noviembre, 1998. Disponible en *www.ed.gov/pubs/TIMSSBrief/student.html* (Recuperado el 6/2/2002).

6. L. S. Grønmo, and R. V. Olsen, "TIMSS versus PISA: The Case of Pure and Applied Mathematics", documento presentado en the Second IEA International Research Conference, Washington, DC, del 8 al 11 de noviembre, 2006.
7. Por ejemplo vea, I. V. Mullis *et al. TIMSS 1999 International Mathematics Report* (Chestnut Hill, MA: International Study Center, Boston College, 2000), cap. 3.
8. Ibíd., Exhibe 1.1 y 1.4.
9. Estos resultados se pueden encontrar en I. V. S. Mullis *et al.*, *TIMMS 2003 International Mathematics Report* (Chestnut Hill, MA: International Study Center, Boston College, 2004), 34.
10. Ibíd., Exhibe 1.1 y 1.2.
11. Organization for Economic Co-operation and Development Directorate of Education, *Learning for Tomorrow's World: First Results from PISA 2003* (Paris: OECD, 2004).
12. Warwick B. Elley, "*How in the World do Children Read?*" The Hague: International Association for the Evaluation of Educational Achievement, 1992.

Capítulo 6. ¿Qué influye en los resultados de las pruebas? (o cómo no escoger una escuela)

1. J. S. Braswell *et al.*, *The Nation's Report Card: Mathematics 2000* (Washington, DC: National Center for Education Statistics, 2001), 153.
2. Comunicado personal, febrero 1996.
3. W. H. Schmidt, "High-School Course-Talking: Its Relationship to Achievement," *Journal of Curriculum Studies* 15, no. 3 (1983): 311-332.
4. Por ejemplo vea P. Kaufman and K.A. Rasinski, *National Education Longitudinal Study of 1988: Quality of the Responses of Eight-Grade Students in NELS: 88* (Washington, DC: U.S. Department of Education, National Center for Education Statistics, 1991).
5. N. Caplan, M. H. Choy, and J. K. Whitmore, "Indochinese Refugee Families and academic Achievement," *Scientific American* (febrero 1992): 36-42; N. Caplan, M. H. Choy, and J. K. Whitmore, *Children of the Boat People: A Study of Educational Success* (Ann Arbor: University of Michigan Press, 1991).
6. B. Hart and T. R. Risley, *Meaningful Differences in the Everyday Experience of Young American Children* (Baltimore: P.H. Brooks, 1995).
7. Para la discusión de esta evidencia, vea D. Koretz, "What Happened to Test Scores, and Why?" *Educational Measurement: Issues and Practice* 11, no. 4 (Invierno 1992): 7-11. Una discusión más extensa puede encontrarse en D. Koretz, *Educational Achievement: Explanations and Implications of Recent Trends* (Washington, DC: Congressional Budget Office, agosto, 1987).
8. Koretz, *Educational Achievement.*

Capítulo 7. Error y confiabilidad:
¿qué tanto no sabemos de lo que estamos hablando?

1. No Child Left Behind Act, 20 USC 6311(b)(3)(C).
2. Massachusetts Department of Education, *Guide to the MCAS for Parents/Guardians* (Malden, MA: Massachusetts Department of Education, n.d.), 9.
3. R. K. Hambleton *et al.*, *Review of the Measurement Quality of the Kentucky Instructional Results Information System*, 1991-1994 (Frankfort: Office of Education Accountability, Kentucky General Assembly, Junio, 1995).
4. College Entrance Examination Board, *SAT Program Handbook 2004-2005* (New York: CEEB, 2005), 38-40.
5. College Entrance Examination Board, *Test Characteristics of the SAT: Reliability, Difficulty Levels, Completion Rates*. Disponible en: *www.collegeboard.com/prod_downloads/highered/ra/sat/sat-test-characteristics.pdf* (Recuperado el 7/24/07).
6. H. D. Hoover *et al.*, *Iowa Test of Basic Skills, Interpretive Guide for School Administrators* (Chicago: Riverside Publishing, 2003).
7. Massachusetts Department of Education, *2003 MCAS Technical Report* (Malden, MA: Massachusetts Department of Education, 2003), 67.
8. *Washington Assessment of Student Learning, Washington Alternate Assessment System (WAAS), 2001 Technical Report* (Itasca, IL, Riverside Publishing Company, n.d.).
9. T. J. Kane, and D. O. Staiger, "The Promise and Pitfalls of Using Imprecise School Accountability Measures," *Journal of Economic Perspectives* 16, no. 4 (2002): 91-114.
10. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing* (Washington, DC: American Educational Research Association, 1999), vea el Estándar 13.7.

Capítulo 8. Informe sobre desempeño: estándares y escalas

1. *This is Spinal Tap*, dirigido por Rob Reiner (1984).
2. M. Sacchetti, "Teachers' Math Skills Are Targeted," *Boston Globe*, 2 de enero, 2007, B1, B3.
3. M. Perie, W. Grigg, and G. Dion, National Center of Education Statistics, *The Nation's Report Card: Mathematics 2005, NCES 2006-453* (Washington, DC: U.S. Government Printing Office, 2005). Vea la p.16.
4. Existen muchos métodos para establecer los estándares. Es difícil y superfluo para el propósito de este capítulo entrar en detalles. Para aquellos interesados en adentrarse en el tema, les recomiendo como inicio una introducción muy clara escrita por Ronald Hambleton, "Setting Performance Standards on Achievement Tests: Meeting the Requirements of Title I," en L. Hansche *et al.*, *Handbook for the Development of Performance Standards: Meeting the Requirements of Title I* (Washington, DC: Council of Chiefs State School Officers, 1998). Disponible en: *www.ccsso.org/content/pdfs/hansche.pdf*.

5. James S. Braswell *et al.*, U.S. Department of Education, *The Nation's Report Card: Mathematics 2000* (Washington, DC: Office of Educational Research and Improvement, 2001), Figura 1.3.
6. W. J. Popham, *Criterion-Referenced Measurement* (Engelwood Cliffs, NJ: Prentice-Hall, 1978); Hambleton, "Setting Performance Standards on Achievement Test," en Hansche, ed. *Handbook for the Development of Performance Standards*, 87-115.
7. R. M. Jaeger, "Certification of Student Competence," en R. L. Linn, ed., *Educational Measurement*, 3rd. ed. (New York: American Council of Education/Macmillan, 1989), 485-514.
8. Por ejemplo , vea R. L. Linn, "Performance Standards: Utility for Different Uses of Assessments," *Education Policy Analysis Archives* 11, no. 3 (2003); Disponible en: *http://epaa.asu.edu/epaa/v11n31/* (Recuperado el 6/30/06; Revista disponible en formato electrónico únicamente y los artículos no tienen número de página dentro de cada ejemplar); y L. A. Shepard, "Implications for Standard Settings of the National Academy of Education Evaluation of the National Assessment of Educational Progress Achievement Levels," en *Proceeding of the Joint Conference on Standard Settings for Large-Scale Assessments*, vol. 2 (Washington, DC: National Assessment Governing Board and National Center for Education Statistics, 1994), 143-160.
9. R. M. Hauser, C. F, Edley Jr., J. A. Koening, and S. W. Elliott, eds., *Measuring Literacy: Performance Level for Adults*, Reporte interno (Washington, DC: National Academies Press, 2005), Tablas 5-5b y 5-5c.
10. Para datos de los estándares estatales, vea L. Olson, "Defying Predictions, State Trends Prove Mixed on Schools Making NCLB Targets," *Education Week*, Septiembre 7, 2005, y tablas asociadas; que están disponibles en: *www.edweek.org/ew/articles/2005/09/07/02ayp.h25.html?qs=_defying_predictions_*. (Recuperado el 1/5/07). Para datos sobre los porcentajes de competencia de los estados en el NAEP, vea Perie, Griggs, and Dion, *The Nation's Report Card.*
11. H. I. Braun and Jiahe Qian, "*An Enhanced Method for Mapping State Standard onto the NAEP Scale*," en N. J. Dorans, M. Pommerich, and P.W. Holland, ed., *Linking and Aligning Scores and Scales* (New York Springer-Verlag, 2007), 313-338.
12. H. D. Hoover, "Some Common Misconceptions about Tests and Testing." *Educational Measurement: Issues and Practice* 22, no.1 (2003): 5-13.
13. E. H. Haertel and D. E. Wiley, "Response to the OEA Panel Report, "Review of the Measurement Quality of the Kentucky Instructional Results Information System, 1991-1994," Documento no publicado preparado para the Kentucky Department of Education and Advanced Systems in Measurement and Evaluation, 1995.
14. Por ejemplo, vea, R. M. Hauser *et al.*, eds., Measuring Literacy: Performance Levels for Adults (Washington, DC: National Academies Press, 2005).

15. R. L. Linn, "Assessment and Accountability", *Educational Researcher* 29, no. 2 (2000): 4-16.
16. Linn, "Performance Standards: Utility for Different Users of Assessments"
17. Por ejemplo, vea, Perie, Grigg, and Dion, *The Nation's Report Card*, 14.
18. College Board, *2005 College-Bound Seniors* (New York: College Board, 2005).

Capítulo 9. Validez

1. J. Macur, "Testing Is Just First Step in Case against Landis," *New York Times*, 5 de agosto, 2006.
2. C. Bradley and E. Posner, "Signing Statements: It's a President's Right," *Boston Globe*, 3 de agosto, 2006, Completo.
3. *House Call with Sanjay Gupta: A look at What Causes Headaches*, 25 de septiembre, 2004. De *http:// transcripts.cnn.com/ TRANSCRIPTS/0409/25/ hesg.00 Html* (Recuperado el 8/21/06).
4. Esta tarea se puede encontrar en: *http:// pals.sri.com/task/5-8/ME127/* (Recuperado el 1/10/07).
5. Por ejemplo, vea L.S. Hamilton, E.M. Nussbaum, and R. E. Snow, "Interview Procedures for Validating Science Assessments," *Applied Measurement in Education* 10 (1997): 181-200.
6. D. Koretz *et al.*, "The Vermont Portfolio Assessment Program: Findings and implications," *Educational Measurement: Issues and Practice* 13, no. 3 (1994): 5-16.
7. R.K. Hambleton *et al.*, *Review of the Measurement Quality of the Kentucky Instructional Results Information System, 1991-1994* (Frankfort: Office of Education Accountability, Kentucky General Assembly, junio, 1995.
8. J. P. Greene, M. A. Winters, and G. Foster, *Testing High States Tests: Can We Believe the Results of Accountability Test?* Civic Report 33, Resumen ejecutivo (New York: The Manhattan Institute, 2003).
9. Hamilton, Nussbaum, and Snow, "Interview Procedures for Validating Science Assessments."

Capítulo 10. Resultados inflados de las pruebas

1. Por ejemplo, vea, R. L. Linn and S.B. Dunbar, "The Nation's Report Card Goes Home: Good News and Bad about Trends in Achievement, "*Phi Delta Kappan 72*, no. 2 (octubre, 1990): 127-133; B. Fuller *et al.*, *Is the No Child Left Behind Act Working? The Reliability of How States Track Achievement* (Berkeley: University of California, Policy Analysis for California Education, 2006).
2. Donald T. Campbell, "Assessing the Impact of Planned Social Change," En G. M. Lyons, ed., *Social Research and Public Policies: The Dartmouth/OECD* Conference (Hanover, NH: Public Affairs Center, Dartmouth College, 1975) 35.

3. L. Zuckerman, "Airline Math, an Early Arrival Doesn't Mean You Won't Be Late," *New York Times*, 26 de diciembre, 2000.
4. Por ejemplo, vea, A. Hickman *et. al.*, "Did Sun Cheat?" *PC Magazine*, 6 de enero, 1997: y P. H. Lewis, "How Fast is Your System? That Depends on the Test," *New York Times*, 10 de septiembre, 1998, EI.
5. J. Markoff, "Chip Maker Takes Issue with a Test for Speed," *New York Times*, 27 de agosto, 2002, C3.
6. J. H. Cushman, "Makers of Diesel Truck Engines Are under Pollution inquiry," *New York Times*, 11 de febrero, 1998.
7. P. Farhi, "Television´s 'Sweeps' Stakes: Season of the Sentational Called a Context Out of Control," *Washington Post*, 17 de noviembre, 1996, A01.
8. M. Santora, "Cardiologists Say Rankings Sway Choices on Surgery." *New York Times*, 11 de enero, 2005.
9. S. P. Klein *et al.*, *What Do Test Scores in Texas Tell Us?* Documento número IP-202 (Santa Monica, CA: Rand, 2000). Disponible en: *www/rand.org/publications/IP/IP202/*.
10. Para contar con un resumen de la investigación más importante sobre la respuesta de maestros y directores a las evaluaciones, vea B. Stecher, "Consequences of Large-Scale, High- Stakes Testing on School and Classroom Practice," en L. Hamilton *et al.*, *Test-Based Accountability: A Guide for Practitioners and Policymakers* (Santa Monica, CA: Rand, 2002). Disponible en: *www.rand.org/Publication/MR/MR1554/MR1554.ch4.pdf.*
11. J. Rubinstein, *Princeton Review: Cracking the MCAS Grade 10 Mathematics* (New York: Random House, 2000) 15.
12. J. Rubinstein, *Princeton Review: Cracking the MCAS Grade 10 Mathematics*, p. 31.
13. V. Strauss, "Review Tests Go Too Far, Critics Say," Washington Post, 10 de julio, 2001, A09.

Capítulo 11. Impacto adverso y sesgo

1. Se pueden encontrar estos datos y evidencias sobre la sólida relación entre el porcentaje que presenta la prueba y la media de calificación estatal en E. B. Page and H. Feifs, "SAT Scores and American States: Seeking for Useful Meaning" *Journal of Educational Measurement* 22, no. 4 (1985): 305-312.
2. Vea *www.ucop.edu/news/factsheets/2006/fall_2006_admissions_table_c.pdf* (Recuperado el 2/6/07).
3. College Board, *2006 College-Bound Seniors* (New York: College Entrance Examination Board, 2006).
4. Vea *www.users.com/usnews/edu/college/ranking/brief/t1natudoc_brief.php* (Recuperado el 3/23/07).

5. Para una revisión sobre las diferencias de género en el desempeño en diversas pruebas, vea Warren W. Willingham and Nancy S. Cole, *Gender and Fair Assessment* (Mahwah, NJ: Lawrence Erlbaum, 1997). Los resultados más recientes del NAEP pueden encontrarse en National Center for Education Statistics, *The Nation's Report Card: Mathematics 2005* (Washington, DC: U.S. Department of Education, 2005).
6. Vea B. Bridgeman, L. McCamley-Jenkins, and N. Ervin, *Prediction of Freshman Grande- Point Average from the Revised and Recentered SAT I: Reasoning Test*, College Board Research Report no. 2000-1, ETS RR No.00-1 (New York: College Entrance Examination Board, 2000).

Capítulo 12. Evaluación de estudiantes con necesidades especiales

1. National Research Council, Committee on Goals 2000 and the Inclusion of Students with Disabilities, Educating *One and All: Students with Disabilities and Standards-Based Reform* (Washington, DC: National Academy Press, 1997) 103.
2. R.C Kessler *et al.*, "Prevalence, Severity, and Comorbidity of Twelve Month DSM-IV Disorders in the National Comorbidity Survey Replication (NCS-R)," *Archives of General Psychology* 63, no. 6 (2005): 617-627.
3. 29 U.S.C.§§794 et seq.; También, the Americans With Disabilities Act (42U.S.C. §§121011 *et seq.*) y the Goals 2000: Educate America Act (20 U.S.C. §§5801 et seq.).
4. 20 U.S.C § 1412(a)(17).
5. 20 U.S.C 6311(b)(3)(C).
6. M. Rogosta and B. Kaplan, "Views of Disabled Students," en W. W. Willingham *et al.*, eds. *Testing Handicapped People* (Boston: Allyn and Bacon, 1988), 57-70.
7. Para una interesante discusión e investigación relacionada sobre este ejemplo, vea *www.mrc-cbu.cam.ac.uk/personal/matt.davis/Cmabrigde/*.
8. W.W. Willingham *et al. Testing Handicapped People* (Boston: Allyn and Bacon, 1988), 129.
9. *Southeastern Community College v. Davis*, 442 U.S 397(1979). Para una discusión de este caso, vea *http://caselaw.lpfindlaw.com/scripts/getcase.pl?court=US&vol=442&invol=397* (Recuperado el 7/21/06).
10. M.L. Thurlow *et al.*, *2001 State Policies on Assessment Participation and Accommodations*, Synthesis Report 46 (Minneapolis: University of Minnesota, National Center on Educational Outcomes, 2002). Disponible en *http://education.umn.edu/NCEO/OnlinePubs/ Synthesis46.html* (Recuperado el 10/21/07).
11. National Research Council, *Educating One and All*, 177-178.

12. Para el aviso de la propuesta reguladora, vea Federal Register, 30 de marzo, 2003, 13796ff. Para la aplicación de los reglamentos finales, vea *Federal Register*, 9 de diciembre 2003, 68698ff. La explicación de la escisión de referencias a los estudiantes cuyo desempeño se encuentra tres desviaciones estándar por debajo de la media se puede encontrar en la p. 68704.
13. 20 U.S.C. §6311(b)(1)(D)(ii)(II).
14. W. Piper, *The Little Engine That Could* (New York: Platt and Munk, 1930).
15. Vea: *Federal Register*, 7 de abril, 2007, 17748ff.
16. Ibíd., 17748.

Capítulo 13. Usos razonables de las pruebas

1. K. W. Areson, "Math and Science Tests Find 4th and 8th Graders" En U.S Still Lag Many Peers," *New York Times*, 15de diciembre, 2004. Disponible en: *www.nytimes.com/2004/12/15/education/15math.html?ex=1175227200&en=c9be2729f4b06bcd&ei= 5070* (Recuperado el 3/27/07).
2. National Center for Education Statistics, "International Comparisons in Education, Trends in International Mathematics and Science Study: Mathematics and Science Achievement of Eighth-Graders in 1999." Disponible en: *http://nces.ed.gov/times/results99_1.asp* (Recuperado el 3/27/07).
3. Kentucky Department of Education, *KIRIS Accountability Cycle 2 Technical Manual* (Frankfort: Kentucky Department of Education, 1997), 14-17.
4. D. T. Campbell, "Assessing the Impact of Planned Social Change," en G. M. Lyons, ed., *Social Research and Public Policies* (Hanover, NH: Public Affairs Center, Dartmouth College,1975), 3-45.

El ABC
de la
evaluación educativa

Reconocimientos

Una tarde, hace algunos años, varios estudiantes que habían llevado uno de mis cursos sobre medición educativa asistieron a una conferencia impartida por uno de los responsables de política educativa más importantes del país. Su presentación enojó a los estudiantes porque aunque muchas de sus afirmaciones acerca de la evaluación educativa eran totalmente erróneas, las presentaba con absoluta certeza y seguridad en sí mismo. Con mucha frecuencia encuentro este tipo de reacción entre los alumnos. Sin embargo, una de ellos estaba especialmente molesta y esa noche me escribió un largo y furioso correo electrónico. Terminó diciéndome que tenía que escribir un libro que ayudara a las personas no especializadas a entender la evaluación educativa. Había estado pensando en eso durante algún tiempo y su nota me hizo tomar la decisión. Para empezar, quiero agradecer a los muchos estudiantes que me exhortaron a escribir este libro, en especial a Marina Lang, que me envió el correo electrónico esa noche, y a Chris Olson Lanier, una antigua editora de adquisiciones que compartió conmigo su experiencia y me puso en movimiento.

Debo agradecer a muchos otros su ayuda en la hechura de esta obra. Deseo expresar mi gratitud a la Carnegie Corporation of New York y, en especial, a Dan Fallon, presidente de la División de Educación de la fundación, por su ayuda para apoyar este trabajo. Estoy en deuda con mi editora en Harvard University Press,

Elizabeth Knoll, quien estuvo a la altura de su excelente reputación y no sólo me orientó de manera siempre útil y perceptiva sino que también me brindó las críticas necesarias con notable tacto. Igual que la mayoría de los académicos, por lo regular no me gusta que nadie juegue con mis textos, pero esperaba con entusiasmo los paquetes con los comentarios de Elizabeth. También le debo mucho a mi esposa, Doreen Koretz, por animarme a escribir el libro y luego sufrir las consecuencias con buen humor y su habitual paciencia. Agradezco a los muchos colegas y amigos, pues sin su entusiasmo habría sido mucho más difícil perseverar para alcanzar las metas de este libro. Por último, como una persona que ha enseñado en casi todos los niveles desde cuarto de primaria hasta el posgrado y que se toma muy en serio la enseñanza, quiero agradecer las enseñanzas que me han brindado tantos colegas. Son muchos para poder mencionarlos a todos, pero hay dos a quienes quiero destacar: H. D. Hoover, profesor emérito de la Universidad de Iowa, y Robert Linn, profesor emérito de la Universidad de Colorado, eruditos destacados y buenos amigos que a lo largo de muchos años han compartido su experiencia con una generosidad indefectible.

Índice temático

Adecuaciones para estudiantes con discapacidades, 333-341, 350-353; para estudiantes con dominio limitado del inglés, 361-363
Acromatopsia, 338-339
ACT, 44, rango de resultados en, 183; escala de, 91, 238
Actitudes, de los examinados, 27
Acuerdo entre jueces, 177
Agudeza visual, 259, 335-336; en la acromatopsia, 338-341
Álgebra, 46-49, 50, 149-150
Alineación, 296-297, 301
Aprendiz del inglés, 80n
Aprobación/reprobación, punto de corte, 187-189
Asociación Internacional para la Evaluación del Logro Educativo (*International Association for the Evaluation of Educational Achievement, IEA*), 124
Azar, 202, 204-205
Báscula de baño, 173-175, 176, 202
Braille, 340, 341
Bush, George W., 11, 15, 252
Calidad del maestro, 138
Calificación de ensayos, 177
Calificación estándar de desarrollo, 247
Calificaciones, 16; sesgo y, 266, 324; en la universidad, 322-324; SAT y, 323-325
Cannell, John, 70
Carrera a pie, 264
Chicos de la burbuja, 227
Chips de computadora, estadísticas del desempeño de, 278
Cirugía cardiaca, calificación de los médicos en, 280
Coeficiente de confiabilidad, 185-187
Coeficiente intelectual (CI), 12
Competencia (o dominio) limitada en el inglés, 80, 106-107, 327, 360-365; adecuaciones para, 361-363; lengua materna y, 363; investigación sobre, 363
Competencia relativa, 34
Competencia, 34-35, 62, 213, 225-227. Véase también Estándares de desempeño
Concurrencia temporal,155, 158-160
Conducta, criterio, 48,51
Confiabilidad entre jueces, 190
Confiabilidad, 37, 171; consistencia de la clasificación y, 186-189; diferencias en las calificaciones de grupo y, 321-322; mejoras en, 190; consistencia interna y, 189, 190; entre jueces, 190; complejidad de la prueba y, 192-193; longitud de la prueba y, 192; validez y, 260-263. Véase también Error de medición
Connerly, Ward, 311
Conocimiento, aplicación del, 46-48
Consecuente, validez, 254-255
Consejo Nacional de Investigación, 11, 18
Consistencia entre jueces, 177, 190
Consistencia interna, 175, 190-191

Consistencia: entre medidas alternativas, 262-263; interna, 175, 190-191
Correlación, 147-148; entre medidas de desempeño múltiples, 262-263, 266-271
Corrupción de las medidas, 277-280
Cronbach, Lee, 254
Curva de campana, 94-95
Darlington, Dick, 148
Datos de impacto, de los estándares de desempeño, 218-219, 229-230
Desigualdades educativas, 164-167
Desviación estándar, 92-97; evaluaciones internacionales y, 132-133
DIF. Véase Funcionamiento diferencial del reactivo
Diferencia promedio: estudiantes afroamericanos frente a estudiantes blancos, 117-118, 167-168; estudiantes asiáticoestadounidenses frente a estudiantes blancos, 120-121; estudiantes hispanos frente a estudiantes blancos no hispanos, 118-120; hombres frente a mujeres, 317-318
Diferencias de grupo, 117-121, 167-168, 307-309
Dificultad del reactivo, 33
Discapacidad auditiva, 347-348
Discapacidad irrelevante para el constructo, 339-340
Discapacidad relevante para el constructo, 344-348, 352-353
Dislexia, 345-347, 353
Distractores, 27
Distribución Gaussiana, 94-95
Distribución normal, 94-95
Dominio, 25-26; muestra del, 27-33, 256
Efecto Berkeley, 310-316
Efecto de la cohorte, 109, 161
Efecto de piso, 353
Efecto del periodo, 109
Efectos de la composición, 104-107, 115-116, 161
Elaboración (construcción) de la prueba, 18-19, 26-27, 191-192, 380-383; ejemplo de, 28-34, 36-37; evaluación internacional y, 124-125; muestreo matricial en, 76-77; inflación de resultados y, 75; reactivos simples frente a reactivos complejos en, 50-52, 58-59
Elección de escuela, 5-6, 378-379
Elección presidencial (2004), 21, 236-237
Elecciones, encuestas durante las, 26-28
Encuestas, 16-19; precisión de las, 17-19; contra evaluación de logro, 24-26; sesgo en, 194-195; margen de error en las, 194; falta de respuesta en, 194-195; preguntas para, 23 ; respuesta en, 23-35, 194-195; muestreo en, 22-23, 194-195
Equivalente de grado, 246-247, 285-287
Equivalentes de la curva normal, 245
Error de medición, 36-37, 171-189; frente a sesgo, 174-175, 309; definición de, 172-175; diferencias en la calificación de grupo y, 320-321; cuantificación del, 180-189; rango probable de resultados y, 171-173; coeficiente de confiabilidad y, 184-187; métodos de calificación y, 177-180; fuentes de, 175-180; subestimación del, 189
Error de muestreo, 170, 189-201; calificaciones desagregadas y, 168-200; tamaño del, 196
Error estándar de medición (EEM), 180-184, 304

Error, 170-171; definición del, 173; de muestreo, 170, 189-201 Véase también Error de medición; Error de muestreo
Escala de puntajes Z, 313
Escala estandarizada, 97-98
Escalamiento, 210
Escalas de desarrollo, 248
Escalas de intervalo, 247-248
Escalas de temperatura, 90-91, 210, 234-235
Escalas, 16-18, 90-100, 233-250; arbitrarias, 91, 238-241; comparaciones de, 91-94; elaboración (o construcción) de, 239-240; de desarrollo, 246-247, 248; de intervalo, 247-248; transformación lineal de, 235; transformación no lineal de, 238-239; normas y, 243-244; rango percentil y, 99, 244; contra estándares de desempeño, 211-233, 241-243, 376-377; desviación estándar de, 92-97; estandarizadas, 97-100; de temperatura, 90, 210, 234-235; tipos de, 210-211
Establecimiento de estándares, 50-51
Estándares de aprendizaje, 55
Estándares de contenido, 77-80, 212
Estándares de desempeño, 77, 78-80, 211-233; método Angoff para, 215-219, 222; naturaleza arbitraria de, 220-226; método del marcador para, 219-220; consistencia de la clasificación y, 187-189; definición de, 78-79, 212; establecimiento de, 213-223; variación entre grados en, 224-225; comparaciones de grupo y, 227-229; consecuencias instruccionales de, 226-227; interpretación de, 376-379; método Angoff modificado para, 215-219; contra pruebas referidas a normas, 211-212, 229-232; contra escala de calificación, 241-243; variación estatal en, 223-224; objetivos para, 82-85; poco realistas, 163-166, 229-231
Estandarización, 29
Estatus socioeconómico (SES), 122-123, 142-143, 151-153
Estudiantes afroamericanos; prohibición de la acción afirmativa y, 310-316; reporte basado en estándares y, 227-229; mejoras en los resultados de las pruebas por, 161-162; diferencias promedio entre blancos y, 117-119, 167-168
Estudiantes asiaticoestadounidenses, 120-121, 312, 315-316
Estudiantes con discapacidades, 330-359; adecuaciones para, 333-341; 351-353; tiempo adicional para, 342-343; barreras irrelevantes para el constructo, para el desempeño de, 339-340; discapacidad relevante para el constructo y, 344-349, 352-353; identificación de, 329-333, 347-348, 350-351; bajo desempeño, 353-359; investigación sobre, 350-353; problema para, ojeadas rápidas, 341-342; directrices estatales sobre, 349-350; requisitos de evaluación para, 333-334
Estudiantes con necesidades especiales, 80-81, 104, 327-365; NCLB y, 86-87; validez y, 271-273. Véase también Estudiantes con discapacidades; Dominio limitado del inglés
Estudiantes de bajo desempeño, 353-359
Estudiantes hispanos, 119-120

Estudio de un idioma extranjero, 149-150
Estudio Internacional de Tendencias en Matemáticas y Ciencia (*Trends in International Mathematics and Science Study, TIMSS*), 60, 123-133, 166-167
Estudio Internacional del Progreso en la Comprensión Lectora *(Progress in International Reading Literacy Study, PIRLS*), 124, 129-130
Evaluación auténtica, 71, 73-74
Evaluación de alto impacto, 56-57, 156; interpretación de, 373-379; validez de 270-272. Véase también Inflación de resultados o calificaciones
Evaluación de competencia mínima, 66-67
Evaluación de la lectura, 92-94, 96-99; dificultad para la decodificación (en la dislexia), 344-347, 352-353; fluidez, 341-342; comparaciones internacionales en, 129-130; evaluación ITBS de, 267-269; evaluación KIRIS, de, 290-292; evaluación NAEP de, 90, 111, 120, 223-224, 290-292; evaluación PIRLS de, 124, 129-130; remedial, 358
Evaluación de opción múltiple, preparación para, 74-75
Evaluación del desempeño, 70-76
Evaluación exigida por el estado, 56, 67, 81-82; estándares de contenido y, 78-79
Evaluación internacional, 123-133; inconsistencias en, 126-128; interpretación de, 369-373; matemáticas, 128-129, 131; grupo normativo y, 125-127; lectura, 129-130; elaboración de la prueba en, 125; variabilidad en, 131-133, 166-167
Evaluación Nacional del Progreso Educativo (*National Assessment of Educational Progress, NAEP*),14-15, 65-66, 91, 111-114; efectos de la composición en, 115-116; evaluaciones de matemáticas mediante, 14-15, 82-83, 111-114, 115-116, 120-121, 122-123, 162-163, 206-207, 217-218, 223-224; NCLB y, 162-163; estándares de desempeño de, 215-219; evaluación de la lectura mediante, 111, 120, 223, 290-292; resultados en, 243-244; resultados a nivel estatal en, 135-136
Evaluación por portafolios, 64, 179-180, 269
Evaluación Texana de las Habilidades Académicas (*Texas Assessment of Academic Skills, TAAS*) 15-16, 292
Evaluaciones por muestreo matricial, 76
Examen de Registro de Posgrado (GRE), 354
Examen de salida, 66
Exámenes Regentes del Estado de Nueva York (*New York State Regents Examination*), 57, 179
Expresiones idiomáticas, 363-364
Falta de significancia estadística, 204
Funcionamiento diferencial del reactivo (DIF), 317-320
GED, 44
Género, 317-319
Gráfica de pantimedia, 205, 231
Grupo normativo, 59-61
Hallazgos de la abuela, 138, 154, 162
Hambleton, Ron, 220
Hipótesis, 145-146
Histograma, 92, 94
Homogeneidad, 190-191
Hoover, H. D., 108, 149, 190

Howard, Jeff, 63
Impacto adverso, 308-309, 310-316; Véase también Sesgo
Inconsistencia, 173-174; Véase también Error de medición; Error de muestreo
Inferencia: absoluta, 39-40; consecuencias de, 56-57, 200. Véase también interpretación del resultado obtenido en la prueba; Validez
Inflación de resultados 41, 69-70, 110, 156-157, 275-302; alineación y, 296-297, 301-302; trampas y, 280-281, 294-295; preparación y, 297-298; presiones de corrupción y, 277-280; estudio empírico de, 283-290; evaluación de alto impacto y, 15-16, 375; evaluación del desempeño y, 75; reasignación y, 295-296, 299; debilitamiento de la muestra y, 281-282; patrón dentado y, 282-283, 285; gravedad de, 293, 385-386; comparación del formato de la prueba y, 290-291; preparación para la prueba y, 293-302; validez y, 289-290
Ingreso, 153
Inmigración, 105, 120, 121, 154-155
Instrucción dirigida por la medición, 68
Instrucción, evaluación del desempeño e, 74-75
Interpretación de las resultados obtenidos en la prueba, 367-386; para la instrucción, 379-380, 383-384; para comparaciones internacionales, 369-373; problemas en, 368-369; para la calidad de la escuela, 378-379; en pruebas exigidas por el estado, 373-378; pruebas incompletas y, 52-53
Jaeger, Richard, 222
Jencks, Christopher, 160-161
Kane, Tom, 196-199
Kennedy, Edward, 85
Kentucky, evaluación de alto impacto en, 375-376; Evaluación KIRIS en, 290-292; evaluación de portafolios en, 64, 179-180
KIRIS. Véase Sistema de Información de Resultados Instruccionales de Kentucky
Lago Wobegon, efecto del, 69-70
Landis, Floyd, 252
Lanzamiento de una moneda, 202-204
Lector con fluidez, 341-342
Leo, efecto de 200-201
Ley «Que ningún niño se quede atrás» (*No Child Left Behind, NCLB*), 11, 69, 85-87; resultados desagregados para, 197-199; evaluación de estudiantes de bajo logro según, 355-359; estándar de competencia de, 34-35, 39-40, 85-87, 187, 212-213, 227; resultados de, 162-163; evaluación de estudiantes con necesidades especiales bajo la, 334; confiabilidad de la prueba y, 171; validez y, 251. Véase también Inflación de resultados o calificaciones
Ley de Campbell, 277, 385
Ley de Educación Primaria y Secundaria (*Elementary and Secondary Educational Act, ESEA*), 65, 85
Ley de rehabilitación (*Rehabilitation Act*) (1973), 348
Ley para la Educación de Individuos con Discapacidades (*Individual with Disabilities Education Act, IDEA*), 330-331, 333-334
Lindquist, E. F., 44-47, 45-46, 49-50, 253

Líneas aéreas, estadísticas de puntualidad de, 277-278
Linn, Robert, 232-233
Listas de ranking, 140, 253, 306
Madaus, George, 156
Margen de error, 16-17, 170, 194
Matemáticas: variaciones en el logro en, 63; aplicación de, 46-48; evaluación California STAR de, 196-197; evaluación ITBS de, 58-59, 267-269; evaluación MCAS de, 32-33; evaluación NAEP de, 14-15, 82-83, 111-114, 115, 120-121, 123-124, 162, 205-207, 217-218, 223-224; evaluación PISA de, 128-129, 132-133, 370-372; discapacidad lectora y, 346-347, 352-353; diseño de la prueba para, 190-191; especificaciones de la prueba para, 26; evaluación TIMSS de, 82-83, 128, 130-131, 132, 369-373
Medición directa, 48-51
Mehrens, Bill, 265
Messick, Sam, 254
Metas educativas, 45; generalidad de, 46-48; evaluación del desempeño y, 74; próximas, 50
Método Angoff, para los estándares de desempeño, 215-219, 222
Método del marcador, para los estándares de desempeño, 219-220
Miller, George, 85
Muestra, 13-14; matriz, 76-77; encuesta, 21-22, 193-195, 256-257; pregunta/reactivo, 25-26, 27-33; inflación de resultados o calificaciones y, 280-281
NAEP. Véase Evaluación Nacional del Progreso Educativo
No significativa, estadística. 204-205, 207
Normas, 59-60
Observación directa, 48-51
Organización para la Cooperación y el Desarrollo Económico (OCDE), 124
Padres, nivel educativo de los, 153
Patrón dentado, 282-283, 285
Personas que no leen con soltura, 342
PET, 361-362
PISA. Véase Programa para la Evaluación Internacional de los Estudiantes (*Programme for International Student Assessment*)
Población estudiantil: rendición de cuentas y, 84-85; homogeneidad social de, 132-133; cambios en el resultado obtenido en la prueba y, 104-108
Polisemia, 364
Popham, Jim, 220
Porcentaje de cambio en el desempeño, 248-249
Precisión, de las pruebas de logro, 25-28; de las encuestas, 22-24, 25-26
Predicción, encuestas y, 25-27
Preguntas. Véase Reactivos
Preliminary SAT/National Merit Scholarship Qualifyng Test (PSAT), 242-243
Preparación para la prueba, 38-39, 40-41, 110, 293-302
Preparación, 297-300
Principio de los Rolling Stones, 381
Principio de muestreo de la evaluación, 27-33
Probabilidad de respuesta, 219-220
Probabilidad, 202-204
Proceso de admisión a la universidad, 53
Programa de Evaluación del Desempeño Escolar de Maryland (*Maryland School Performance Assessment Program, MSPAP*), 55

Programa para la Evaluación Internacional de los Estudiantes (*Programme for International Student Assessment, PISA*), 124-133; evaluación de matemáticas mediante, 128-129, 132, 370-372
Programación de la televisión, semana de rastreo y, 279
Promedio académico en el primer año en la universidad, 322-323
Promoción social, 254
Propuesta, 243-244, 310-311
Prueba (Evaluación) de Aptitud Escolar. Véase SAT
Prueba adaptativa computarizada, 354
Prueba basada en estándares (referida a estándares), 77-80
Prueba de auditoría, 290
Prueba de vocabulario, 28-33, 35-36, 38-39; preparación para, 39, 40-41
Prueba Nacional de Cualificación del Mérito Académico (*National Merit Scholarship Qualifying Test*) 44
Prueba piloto, 31
Prueba referida a criterio (PRC), 67-68
Prueba referida a normas, 59-63; evaluación internacional y, 60-61, 126-127; contra estándares de desempeño, 211-212, 229-232. Véase también Escalas
Pruebas de Desarrollo Educativo de Iowa (*Iowa Tests of Educational Development*) 44
Pruebas de emisiones de diésel, 279
Pruebas de Habilidades Básicas de Iowa (*Iowa Tests of Basic Skills, ITBS*), 44, 46, 57, 58, 64; correlaciones entre, 267-269; confiabilidad de, 186
Pruebas de logro: complicaciones de, 12-19; referidas a criterio, 67-68; propósito diagnóstico de, 57; discrepancias entre, 15-17; historia de, 57-58; lo incompleto de, 13-14, 45-46, 52-53; resultados inconsistentes de, 14-15, 36-37; variación individual en, 16; limitaciones de, 43; competencia mínima, 66-68; referida a normas, 59-63; basadas en el desempeño, 64-65, 71-76; frente a las encuestas, 24-25; reportes de, 17-18; basadas en estándares, 77-80; usos de, 11-12, 46, 56-57, 68, 253-254. Véase también Sesgo; Interpretación de los resultados obtenidos en la prueba
Pruebas estandarizadas, 29, 51, 63-64
Puntajes crudos, 233
Puntajes estándar, 245
Puntajes T, 245
Punto de corte, 35, 184, 208; establecido en la media de calificación, 313; aprobatoria/reprobatoria, 187-189
Rango de resultados, 171-172
Rango percentil, 92, 99, 244
Reactivo que no discrimina, 34
Reactivos que discriminan, 33-34
Reactivos, 26; sesgo en, 272; complejos, 52, 192-193, 257-259, 261-262; funcionamiento diferencial de, 317-320; dificultad de, 33; que discriminan, 33-34; que no discriminan, 34; selección de, 30-33; especificidad de, 50, 51, 58-59; redacción de, 27
Reasignación, 295-296, 299
Redacción, evaluación directa de, 257
Regulaciones y reseñas de la reglamentación propuesta (RRP), 355-359
Reidy, Ed, 290

Rendición de cuentas, 67, 68-69; comparación de cohortes para, 82; expectativas de desempeño para, 82-84; seguimiento del desempeño para, 81-82; características de la población estudiantil y, 84-85; definición temporal de las expectativas de desempeño para, 83-84
Resultados de aprendizaje, 55
Resultados desagregados de las pruebas, 197-199
Resultados obtenidos en la prueba: disminución de, 100-104, 158-163; distribución de, 92, 94-96, 164; explicaciones para, 144-157; diferencias de grupo en, 117-122, 167-168, 306-309; inconsistencia (incongruencia) en, 16-17, 36-37; bajos, 143-144; factores no educativos y, 137-142, 160; patrones en, 157-168, 282-283, 285; (puntajes) crudos, 17-18, 233; únicos, 208; estatus socioeconómico y, 122-123, 142-143, 151-152. Véase también Error de medición; Confiabilidad; Error de muestreo; Escalas; Validez
Resultados totales agregados, 117, 194-201
Resultados. Véase Resultados obtenidos en la prueba
SAT, desempeño en la universidad y, 322-326; disminución de los resultados en, 100-108; resultados de los estudiantes con discapacidades en, 343; diferencias de grupo en, 306-309; parte de matemáticas de, 115; rango de resultados en, 183-184; coeficiente de confiabilidad de, 185-186; escala de, 91, 238, 239-240, 243, 244-245; inconsistencias en los resultados en, 16-17, 36-37; puntuación de 700 en, 225; promedios estatales en, 140, 306, factores de los estudiantes en, 16-17, 36-37; parte verbal de, 104, 107-108, 114-115
Seguimiento del desempeño, 81-82
Servicio Postal de Estados Unidos: estadísticas de entrega del, 278
Sesgo de deseabilidad social, 23-24
Sesgo de selectividad, 150
Sesgo racial, 24
Sesgo, 18, 173-174, 303-326; calificaciones en la universidad y, 325; factores culturales y, 303-305; definición de, 305; comparación externa y, 321-324; calificación y, 265-266, 324; diferencias en los resultados de grupo y, 307-308, 320-321, 325-326; frente al error de medición, 174-175, 309-310; en las encuestas, 194-195; racial, 24; selectividad, 150; deseabilidad social, 23-24; contenido de la prueba y, 316-319. Véase también Inflación de los resultados
Significancia estadística, 201-207
Sistema de Evaluación Integral de Massachusetts (*Massachusetts Comprehensive Assessment System, MCAS*), 32, 77, 171-172, 176, 183, 186, 232
Sistema de Información de Resultados Instruccionales de Kentucky (*Kentucky Instructional Results Information System, KIRIS*) 290-292
Staiger, Douglas, 196-199
Stewart, Don, 6
Sub-representación del constructo, 256-257

TAAS. Véase Evaluación Texana de las Habilidades Académicas (*Texas Assessment of Academic Skills*)
Tareas, 147-148
This is Spinal Tap, 209-210
TIMSS. Véase Estudio Internacional de Tendencias en Matemáticas y Ciencia (*Trends in International Mathematics and Science Study*)
Título I del Sistema de Evaluación y Reporte (*Title I Evaluation and Reporting System, TIERS*), 65
Trabajo del curso, 149-150
Trampas, 294-295
Triesman, Uri, 324
Una nación en riesgo (*A Nation at Risk*), 69, 101-102
Universidad de California en Berkeley, 310-316
Validez aparente, 263-265
Validez de fe, 263
Validez, 38, 251-273; medidas alternativas y, 266-272; sesgo y, 305; consecuente, 254-255; varianza irrelevante para el constructo y, 257-259; sub-representación del constructo y, 256-257; evidencia convergente de, 267-268; evidencia relacionada con el criterio 265-267; definición de, 252-256, 255-256; evidencia discriminante de, 267-268; aparente, 263-265; de fe, 263; evaluación de alto impacto y, 270-272; confiabilidad y, 260-263; inflación de resultados y, 289-290; estudiantes con necesidades especiales y, 272-273; análisis estadístico y, 272; respuesta del estudiante y, 272-273; contenido de la prueba y, 263-265. Véase también Inflación de los resultados
Variabilidad: explicación de, 163-168; entre grupos, 117-122, 167-168, 307-309: intragrupo, 120-122, 168; evaluaciones internacionales y, 131-133, 166-167; evaluación referida a normas y, 62-63; al calificar, 177-180; estatus socioeconómico y, 122-123
Varianza irrelevante para el constructo, 257-259
Vermont, evaluación de portafolios en, 179-180
Vocabulario, adquisición, 154-155; tamaño del, 29
Vos Savant, Marilyn, 12

El Centro Nacional de Evaluación para la Educación Superior es una asociación civil sin fines de lucro constituida formalmente el 28 de abril de 1994, como consta en la escritura pública número 87036 pasada ante la fe del notario 49 del Distrito Federal. Sus órganos de gobierno son la Asamblea General, el Consejo Directivo y la Dirección General. Su máxima autoridad es la Asamblea General, cuya integración se presenta a continuación, según el sector al que pertenecen los asociados, así como los porcentajes que les corresponden en la toma de decisiones:

Asociaciones e instituciones educativas (40%): Asociación Nacional de Universidades e Instituciones de Educación Superior, A.C. (ANUIES); Federación de Instituciones Mexicanas Particulares de Educación Superior, A.C. (FIMPES); Instituto Politécnico Nacional (IPN); Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM); Universidad Autónoma del Estado de México (UAEM); Universidad Autónoma de San Luis Potosí (UASLP); Universidad Autónoma de Yucatán (UADY); Universidad Nacional Autónoma de México (UNAM); Universidad Popular Autónoma del Estado de Puebla (UPAEP); Universidad Tecnológica de México (UNITEC).

Asociaciones y colegios de profesionales (20%): Barra Mexicana Colegio de Abogados, A.C.; Colegio Nacional de Actuarios, A.C.; Colegio Nacional de Psicólogos, A.C.; Federación de Colegios y Asociaciones de Médicos Veterinarios y Zootecnistas de México, A.C.; Instituto Mexicano de Contadores Públicos, A.C.

Organizaciones productivas y sociales (20%): Academia de Ingeniería, A.C.; Academia Mexicana de Ciencias, A.C.; Academia Nacional de Medicina, A.C.; Fundación ICA, A.C.

Autoridades educativas gubernamentales (20%): Secretaría de Educación Pública.

- Ceneval, A.C.®, EXANI-I®, EXANI-II® son marcas registradas ante la Secretaría de Comercio y Fomento Industrial con el número 478968 del 29 de julio de 1994. EGEL®, con el número 628837 del 1 de julio de 1999, y EXANI-III®, con el número 628839 del 1 de julio de 1999.
- Inscrito en el Registro Nacional de Instituciones Científicas y Tecnológicas del Consejo Nacional de Ciencia y Tecnología con el número 506 desde el 10 de marzo de 1995.
- Organismo Certificador acreditado por el Consejo de Normalización y Certificación de Competencia Laboral (CONOCER) (1998).
- Miembro de la International Association for Educational Assessment.
- Miembro de la European Association of Institutional Research.
- Miembro del Consortium for North American Higher Education Collaboration.
- Miembro del Institutional Management for Higher Education de la OCDE.

La publicación de esta obra la realizó
el Centro Nacional de Evaluación para la Educación Superior, A.C.
Se terminó de imprimir el 30 de septiembre de 2010
en los talleres de Offset Rebosán, S.A. de C.V.,
Av. Acueducto 115, Col. Huipulco Tlalpan, C.P. 14370, México, D.F.
con un tiraje de 500 ejemplares

De manera clara, amena y extensamente documentada, el doctor Daniel Koretz (miembro de la Academia Nacional de Educación de Estados Unidos y profesor de la Escuela de Posgrado en Educación de la Universidad de Harvard) guiará al lector en el conocimiento de las pruebas educativas. Con la lectura de este libro el público no experto podrá adentrarse de una forma sencilla en la complejidad que implica la elaboración de una prueba, mientras que los especialistas conocerán lo intrincado que resulta comunicar los alcances y resultados de los exámenes, de manera simple y entendible, a los usuarios de los servicios educativos y a los responsables de trazar las políticas en materia de educación.

El Ceneval promueve la calidad de la educación mediante evaluaciones válidas, confiables y pertinentes de los aprendizajes. Así contribuye a la toma de decisiones fundamentadas. Con la publicación de esta obra, el Centro continúa atendiendo uno de sus propósitos esenciales: impulsar la cultura de la evaluación en nuestro país.

Rafael Vidal
Director general
Ceneval, A.C.